EE325 Ch.6 Multicollinearity

Read Guajarati Ch. 10

Outline

Nature of Multicollinearity

Multiple Regression

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

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

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \ldots + \lambda_k X_{ki} = 0$$

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

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

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \ldots + \lambda_k X_{ki} + \nu_i = 0$$

where ν_i is a stochastic error term.

The difference between perfect and less than perfect multicollinearity:

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

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

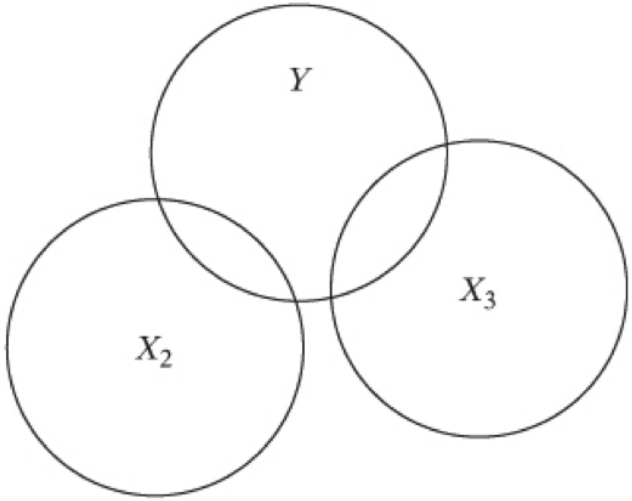
if $\lambda_2 \neq 0$

Perfect Collinearity:

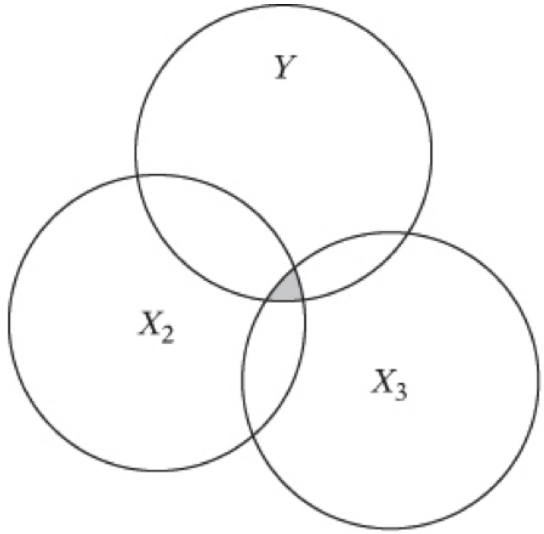
X_2	X_3
10	50
15	75
18	90
24	120
30	150

Imperfect Collinearity:

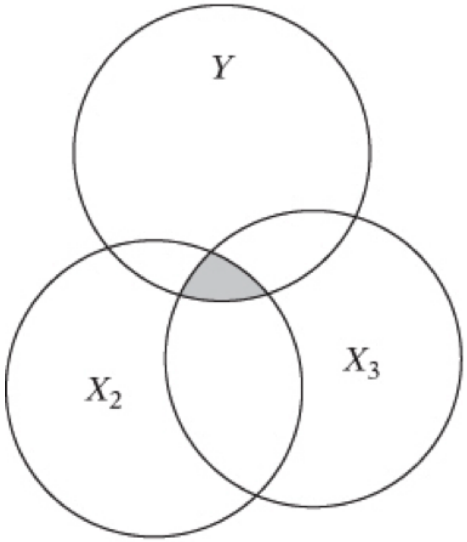
X_2	X_3	ν_i
10	52	2
15	75	0
18	97	7
24	129	9
30	152	2



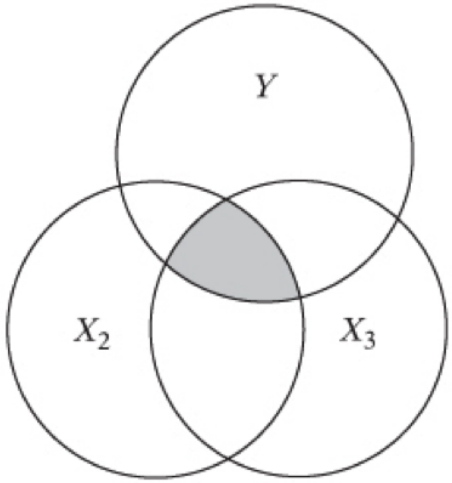
(a) No collinearity



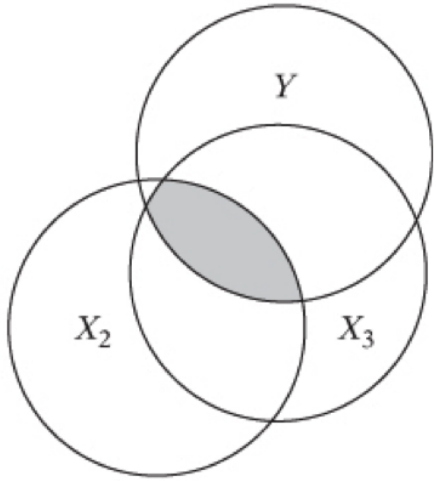
(b) Low collinearity



(c) Moderate collinearity



(d) High collinearity



(e) Very high collinearity

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}^2 + \beta_3 X_{3i}^3 + u_i$$

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

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

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

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

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

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

Consequence of Multicollinearity

Normal OLS estimation without multicollinearity problem:

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

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

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

where
$y_i = Y_i - \bar{Y}$,
$x_{2i} = X_{2i} - \bar{X}_{2i}$,
$x_{3i} = X_{3i} - \bar{X}_{3i}$

Normal OLS estimation with perfect multicollinearity:

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

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

$$\hat{\beta}_3 = \frac{\lambda^2[(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{2i})(\sum x_{3i} x_{3i})]}{\lambda^2[(\sum x_{2i}^2)(\sum x_{2i}^2) - (\sum x_{2i})^2]} = \frac{0}{0}$$

Assume that

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

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

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

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

In case of perfect multicollinearity,

- one cannot get a unique solution for the individual regression coefficients, and
- the variances and standard errors of β_2 and β_3 individually are infinite

Normal OLS estimation with imperfect multicollinearity:

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

Assume that

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

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

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

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

where $\lambda \neq 0$ and $\sum x_i \nu_i = 0$

1. 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.

2. 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

3. 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.

Example:

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

- 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.

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

$$var(\hat{\beta}_2) =$$

$$var(\hat{\beta}_3) =$$

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

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

Variance-Inflating Factor (VIF) shows how the variance of an estimator is inflated by the presence of multicollinearity

$$VIF =$$

when $r_{23} \to 1$, then $VIF \to \infty$

$$var(\hat{\beta}_2) =$$

$$var(\hat{\beta}_3) =$$

TABLE 10.1
The Effect of Increasing r_{23} on var $(\hat{\beta}_2)$ and cov $(\hat{\beta}_2, \hat{\beta}_3)$

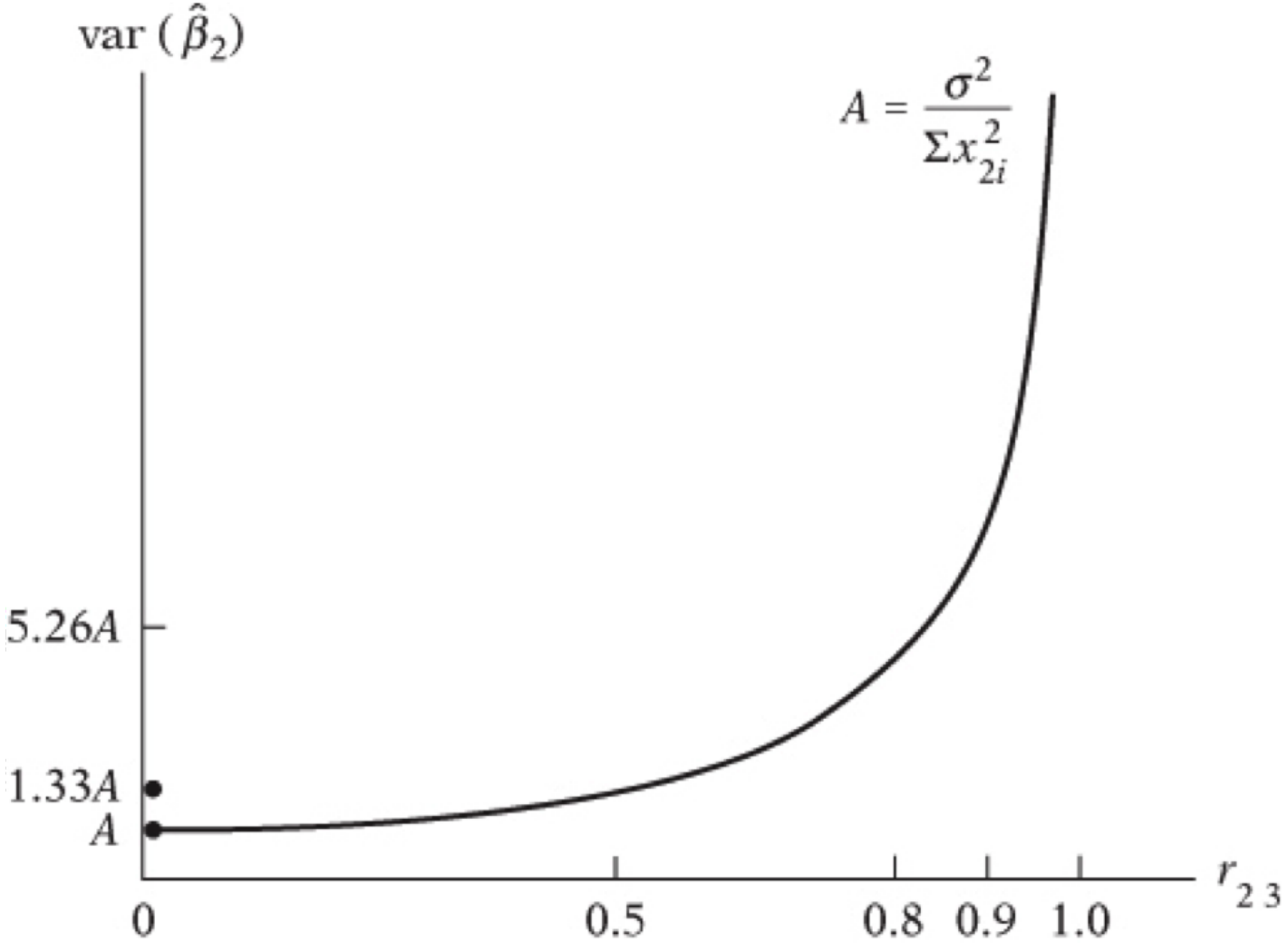
Value of r_{23} (1)	VIF (2)	var $(\hat{\beta}_2)$ (3)*	$\frac{\text{var}\,(\hat{\beta}_2)\,(r_{23} \neq 0)}{\text{var}\,(\hat{\beta}_2)\,(r_{23} = 0)}$ (4)	cov $(\hat{\beta}_2, \hat{\beta}_3)$ (5)
0.00	1.00	$\frac{\sigma^2}{\sum x_{2i}^2} = A$	—	0
0.50	1.33	$1.33 \times A$	1.33	$0.67 \times B$
0.70	1.96	$1.96 \times A$	1.96	$1.37 \times B$
0.80	2.78	$2.78 \times A$	2.78	$2.22 \times B$
0.90	5.76	$5.26 \times A$	5.26	$4.73 \times B$
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	$499.50 \times B$

Note: $A = \dfrac{\sigma^2}{\sum x_{2i}^2}$

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

$\times$ = times

*To find out the effect of increasing r_{23} on 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.


var ($\hat{\beta}_2$)
$A = \frac{\sigma^2}{\Sigma x_{2i}^2}$
5.26A
1.33A
A
0
0.5
0.8
0.9
1.0
r_{23}

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

$var(\hat{\beta}_j)$ is small or large, depending on

1.
2.
3.

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_{23}	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_{23} values are obtained from Table 10.1.

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

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

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

- 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
- Yet the R2 may be so high and leads to rejection of F test

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

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

```
      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  2,     2) =    4.27
       Model |  8.10121951     2  4.05060976           Prob > F      =  0.1899
    Residual |  1.89878049     2  .949390244           R-squared     =  0.8101
-------------+------------------------------           Adj R-squared =  0.6202
       Total |        10       4         2.5           Root MSE      =  .97437

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .4463415   .1848104     2.42   0.137    -.3488336    1.241517
          x3 |   .0030488   .0850659     0.04   0.975    -.3629602    .3690578
       _cons |   1.193902   .7736789     1.54   0.263    -2.134969    4.522774
------------------------------------------------------------------------------
```

$$
\begin{array}{rcccccc}
\hat{Y}_i & = & 1.1939 & + & 0.4463X_{2i} & + & 0.0030X_{3i} \\
se & & (0.7737) & & (0.1848) & & (0.0851) \\
t & & (1.5431) & & (2.4151) & & (0.0358) \\
R^2 & = & 0.8101 & & & & \\
r_{23} & = & 0.5523 & & & & \\
cov(\hat{\beta}_2, \hat{\beta}_3) & = & -0.0087 & & & & \\
df & = & 5 - 3 = 2 & & & &
\end{array}
$$

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

```
      Source |       SS       df       MS              Number of obs =         5
-------------+------------------------------           F(  2,     2) =      4.39
       Model |  8.14324324     2  4.07162162           Prob > F      =    0.1857
    Residual |  1.85675676     2   .928378378          R-squared     =    0.8143
-------------+------------------------------           Adj R-squared =    0.6286
       Total |        10       4         2.5          Root MSE      =    .96352

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .4013514    .272065     1.48   0.278    -.7692498    1.571953
          x3 |    .027027   .1252281     0.22   0.849    -.5117858    .5658399
       _cons |   1.210811   .7480215     1.62   0.247    -2.007666    4.429288
------------------------------------------------------------------------------
```

$$
\begin{aligned}
\hat{Y}_i &= 1.2108 + 0.4014X_{2i} + 0.0270X_{3i} \\
se &\quad (0.7480) \quad (0.2721) \quad (0.1252) \\
t &\quad (1.6187) \quad (1.4752) \quad (0.2158) \\
R^2 &= 0.8143 \\
r_{23} &= 0.8285 \\
cov(\hat{\beta}_2, \hat{\beta}_3) &= -0.0282 \\
df &= 5 - 3 = 2
\end{aligned}
$$

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

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

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

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, but estimators are still BLUE

Detection of Multicollinearity

Kmenta (1986)

- 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
- Since multicollinearity refers to the condition of the explanatory variables that are assumed to be nonstochastic, it is a feature of the sample an not of the population

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

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

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.

Example:

Source	SS	df	MS
Model	8565.55407	2	4282.77704
Residual	324.445926	7	46.349418
Total	8890	9	987.777778

Number of obs	10
F(2, 7)	92.40
Prob > F	0.0000
R-squared	0.9635
Adj R-squared	0.9531
Root MSE	6.808

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x2	.9415373	.8228983	1.14	0.290	-1.004308	2.887383
x3	-.0424345	.0806645	-0.53	0.615	-.2331757	.1483067
_cons	24.77473	6.7525	3.67	0.008	8.807609	40.74186

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

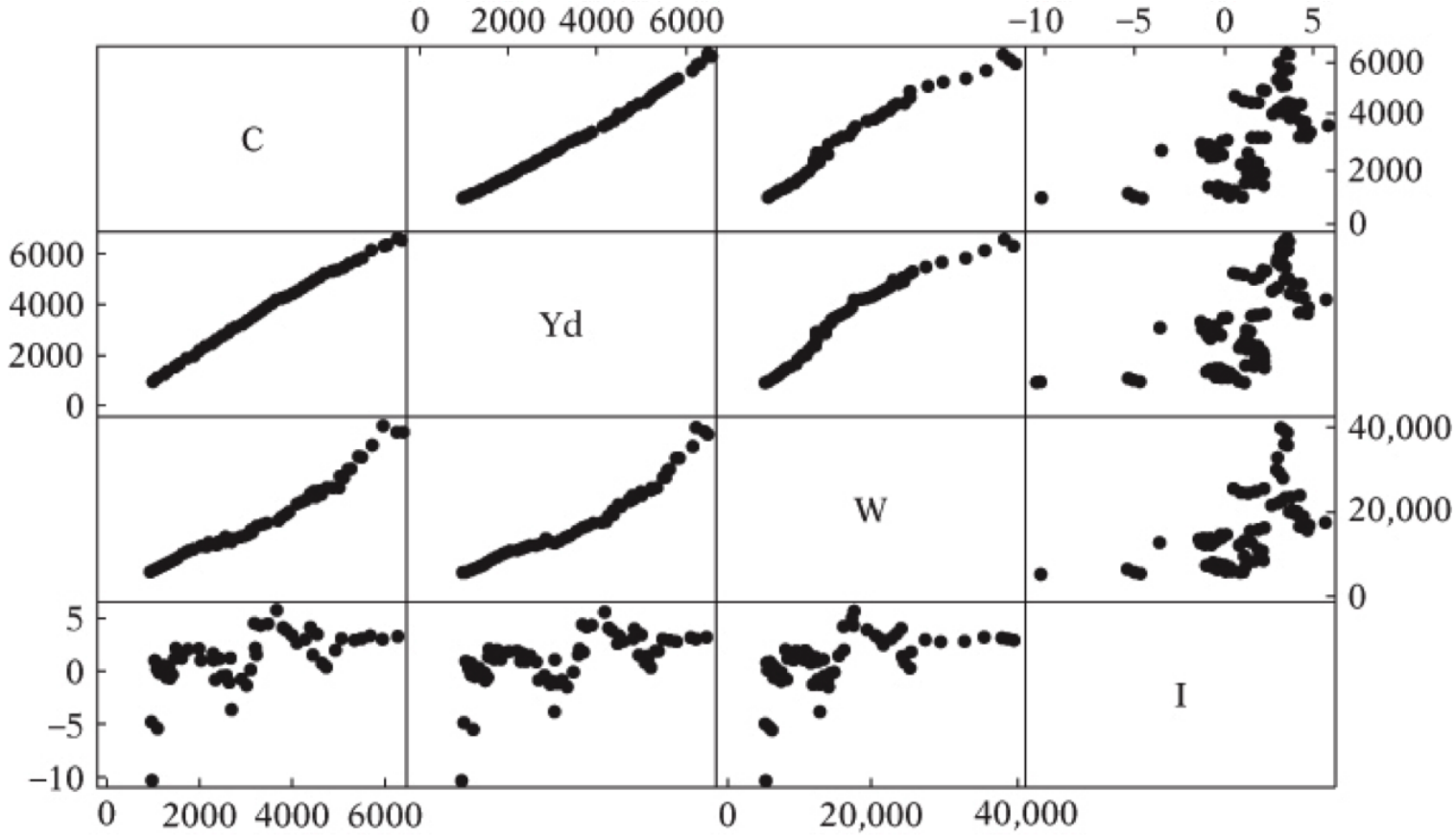
Example: Correlation between income (X_2) and wealth (X_3)

	x2	x3
x2	1.0000	
x3	0.9990	1.0000

Variance-inflating factor (VIF)

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

Scatter Plot



C = Consumption, Y_d = Real disposable personal income,
W = Real wealth, and I = Real interest rate.

Remedial Measures

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

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

A priori information

Given Y_i = consumption, X_{2i} = income, and X_{3i} = wealth,

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \\ \beta_3 &= 0.10\beta_2 \end{aligned}$$

Then

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

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.

Adding or new data

As the sample size increases, $\sum(X_2 - \bar{X}_2)^2$ will generally increases. 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.

- First difference form
- Ratio transformation

First difference form

$$
\begin{aligned}
Y_t &= \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \\
Y_{t-1} &= \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1}
\end{aligned}
$$

- 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 be appropriate in cross-sectional data where there is no logical ordering of the observations

Ratio Transformation

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

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

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

Consumption Expenditure in Relation to Income and Wealth

TABLE 10.5 Hypothetical Data on Consumption Expenditure Y, Income X_2, and Wealth X_3

Y, $	X_2, $	X_3, $
70	80	810
65	100	1009
90	120	1273
95	140	1425
110	160	1633
115	180	1876
120	200	2052
140	220	2201
155	240	2435
150	260	2686

```
      Source |       SS       df       MS
-------------+------------------------------
       Model |  8565.55407     2  4282.77704
    Residual |  324.445926     7  46.349418
-------------+------------------------------
       Total |       8890     9  987.777778

Number of obs =        10
F(  2,     7) =     92.40
Prob > F      =    0.0000
R-squared     =    0.9635
Adj R-squared =    0.9531
Root MSE      =     6.808

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .9415373   .8228983     1.14   0.290    -1.004308    2.887383
          x3 |  -.0424345   .0806645    -0.53   0.615    -.2331757    .1483067
       _cons |   24.77473     6.7525     3.67   0.008     8.807609    40.74186
------------------------------------------------------------------------------
```

$$
\begin{aligned}
\hat{Y}_i &= 24.7747 + 0.9415X_{2i} - 0.0424X_{3i} \\
se &\quad (6.7525) \quad (0.8229) \quad (0.0807) \\
t &\quad (3.6690) \quad (1.1442) \quad (-0.5261) \\
R^2 &= 0.9635 \\
\bar{R}^2 &= 0.9531 \\
df &= 10 - 3 = 7
\end{aligned}
$$

Regression shows that income and wealth together explain about 96% of the variation in consumption expenditure, and yet neither of the slope coefficients is individually statistically significant. Moreover, not only is the wealth variable statistically insignificant but also it has the wrong sign

TABLE 10.6
ANOVA Table for the Consumption–Income–Wealth Example

Source of Variation	SS	df	MSS
Due to regression	8,565.5541	2	4,282.7770
Due to residual	324.4459	7	46.3494

H_o :
H_a :
$F =$
Reject null hypothesis as calculated is greater than critical F

This example shows dramatically what multicollinearity does. The fact that the F test is significant but the t values of X_2 and X_3 are individually insignificant means that the two variables are so highly correlated that it is impossible to isolate the individual impact of either income and wealth on consumption

Source	SS	df	MS
Model	3427202.73	1	3427202.73
Residual	7123.27273	8	890.409091
Total	3434326	9	381591.778

Number of obs	10
F(1, 8)	3849.02
Prob > F	0.0000
R-squared	0.9979
Adj R-squared	0.9977
Root MSE	29.84

x3	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x2	10.19091	.1642623	62.04	0.000	9.81212	10.5697
_cons	7.545455	29.47581	0.26	0.804	-60.42589	75.5168

Source	SS	df	MS
Model	8552.72727	1	8552.72727
Residual	337.272727	8	42.1590909
Total	8890	9	987.777778

Number of obs	= 10
F(1, 8)	= 202.87
Prob > F	= 0.0000
R-squared	= 0.9621
Adj R-squared	= 0.9573
Root MSE	= 6.493

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x2	.5090909	.0357428	14.24	0.000	.4266678	.591514
_cons	24.45455	6.413817	3.81	0.005	9.664256	39.24483

Source	SS	df	MS
Model	8504.87666	1	8504.87666
Residual	385.123344	8	48.1404181
Total	8890	9	987.777778

Number of obs	10
F(1, 8)	176.67
Prob > F	0.0000
R-squared	0.9567
Adj R-squared	0.9513
Root MSE	6.9383

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x3	.0497638	.003744	13.29	0.000	.0411301	.0583974
_cons	24.41104	6.874097	3.55	0.007	8.559349	40.26274

Gujarati, D.N. (2009) Basic Econometrics. 5th ed. Singapore, McGraw-Hill.