part I

Multicollinearity

EE325

Refer to Lecture note before the midterm or Ch 3 in Gujarati (2009).

Assumptions

- Linear regression model, or linear in the parameters
- Fixed X values or X values independent of the error term. We require zero covariance between and each X variables

$$\text{cov}(u_i, X_{2i}) = \text{cov}(u_i, X_{3i}) = 0$$

- Zero mean value of disturbance u_i

$$E(u_i \mid X_{2i}, X_{3i}) = 0, \; for \; each \; i$$

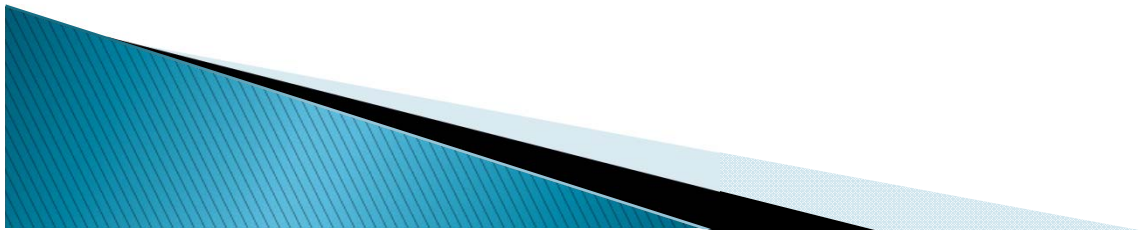
- Homoscedasticity or constant variance of u_i

$$\text{var}(u_i) = \sigma^2$$

- No autocorrelation, or serial correlation, between the disturbances

$$\text{cov}(u_i, u_j) = 0, \quad i \neq j$$

- The numbers of observations n must be greater than the number of parameters to be estimated
- There must be variation in the values of the X variables
- **No exact collinearity between the X variables**
- **There is no specification bias**



What is the nature of multicollinearity?

Is multicollinearity really a problem?

What is its practical consequences?

How does one detect it?

Remedial measures.

The nature of Multicollinearity

Multiple regression

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

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

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \ldots + \lambda_k X_{ki} = 0 \quad (2A)$$

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:

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \ldots + \lambda_k X_{ki} + \nu_i = 0 \quad (2B)$$

where ν_i is a stochastic error term

The difference between perfect and less than perfect multicollinearity

2A

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

2B

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

$$if\ \lambda_2 \neq 0$$

E.g. 2A

Perfect Multicollinearity

$$-5X_{2i} + X_{3i} = 0$$

$X_{3i} = 5X_{2i}$

i	X_2	X_3
1	10	50
2	15	75
3	18	90
4	24	120
5	30	150

E.g 2B

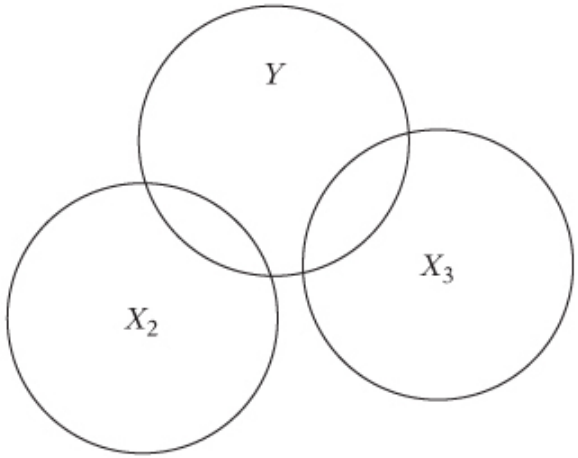
Not Perfect Multicollinearity

i	X_{2i}	X_{3i}	ν_i
1	10	52	-2
2	15	75	0
3	18	97	7
4	24	129	9
5	30	152	2

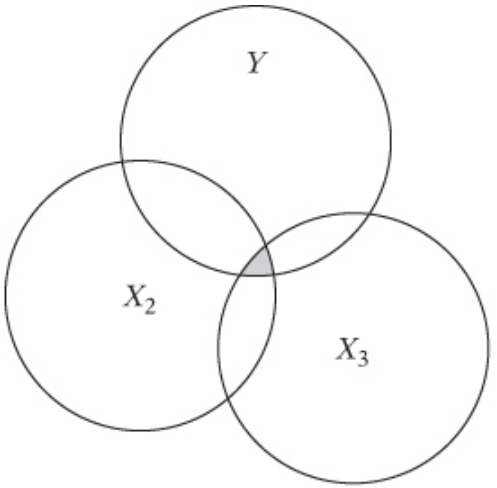
(10) 52 = 0

$$-5X_{2i} + X_{3i} + \nu_i = 0$$

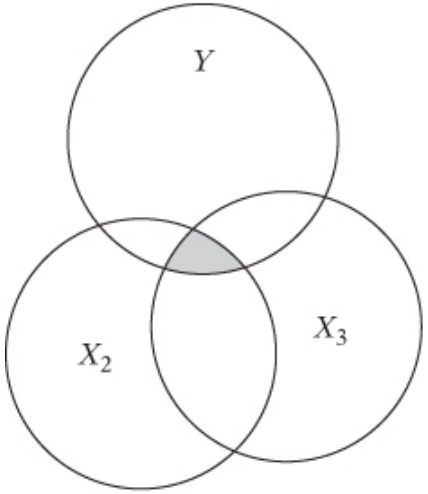
$X_{3i} = 5X_{2i} - \nu_i$



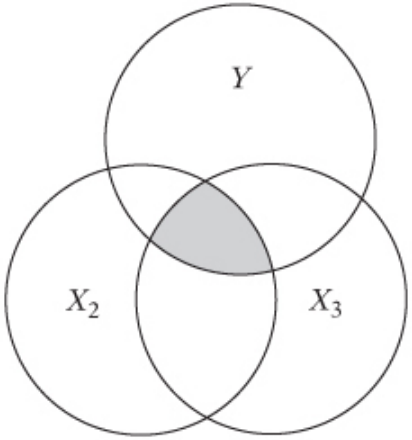
(*a*) No collinearity



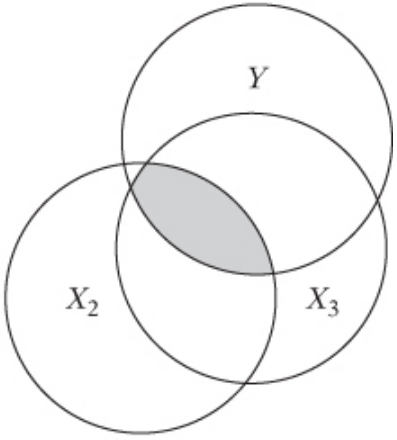
(*b*) Low collinearity



(*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.

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

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

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

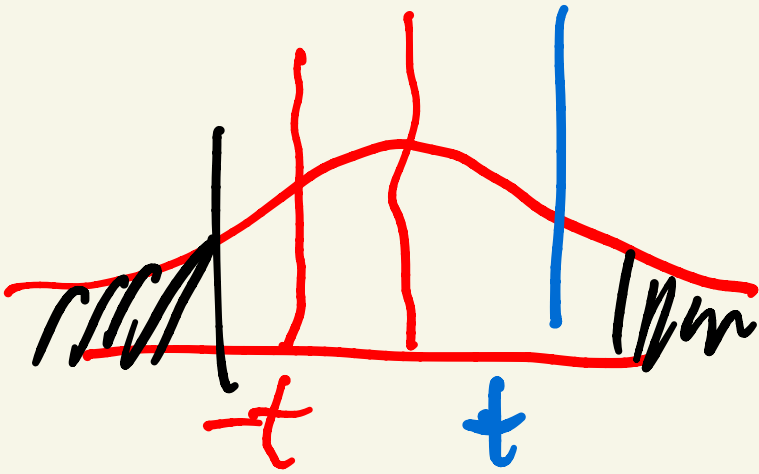
$$X_{2i} = \frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki}$$

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

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

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$



$$t = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$$

When $se(\hat{\beta}_i)$ is large due to multicollinearity

t will be small → less likely to reject H_0.

Several sources of multicollinearity

- The data collection method employed
- Constraints on the model or in the population being sampled

E.g. $Consumption_i = \beta_1 + \beta_2 income + \beta_3 wealth_i$

- Model specification
- An overdetermined model

The data collection method employed

E.g. sampling over a limit range of the values taken by the regressors in the population.

Constraints on the model or in the population being sampled. E.g.

$$\text{Electricity Consumption}_i = \beta_1 + \beta_2 \text{income}_i + \beta_3 \text{house size} + u_i$$

Model specification

E.g. Adding polynomial terms to a regression model, especially when the range of the X variable is small.

$$w_i = \beta_1 + \beta_2 s_i + \beta_3 s_i^2 + \beta_4 E_i + \beta_5 E_i^2$$

An overdetermined model

More explanatory variables than the number of observations

E.g medical research

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

Estimation in the presence of perfect multicollinearity

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

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

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \qquad (1)$$

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

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

Assume that

$$X_{3i} = \lambda X_{2i} \quad (1)$$

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

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

$$x_{3i} = \lambda x_{2i} \quad (4)$$

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

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

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

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

$$\text{or } \text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \quad \text{when } r_{23}^2 = 1 \Rightarrow \text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (0)}$$

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

Estimation in the presence of "High" but "Imperfect" Multicollinearity

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

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

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

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

where $\lambda \neq 0$ and v_i is a stochastic error term such that $\sum x_i \nu_i = 0$

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

" BLUE "

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

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

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.

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

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

Practical Consequences of Multicollinearity

$$t = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$$

$$\hat{\beta}_i \pm t_{\alpha/2}\, se(\hat{\beta}_i)$$

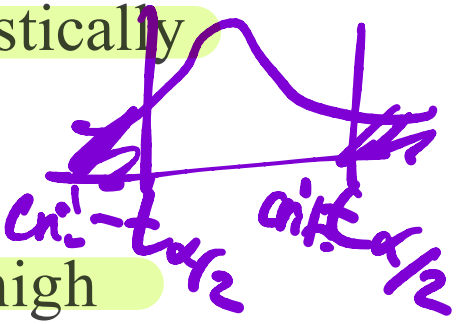
$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

Practical Consequences of Multicollinearity

$se(\hat{\beta}_i) = \sqrt{var(\hat{\beta}_i)}$

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



①

OLS estimators have large variance and covariance

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

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

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

$0 \leq r_{23}^2 \leq 1$

eg. 0.90

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

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

Variance-Inflating Factor (VIF)

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

e.g. $r_{23}^2 = 0.90$

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

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

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

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

Example

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

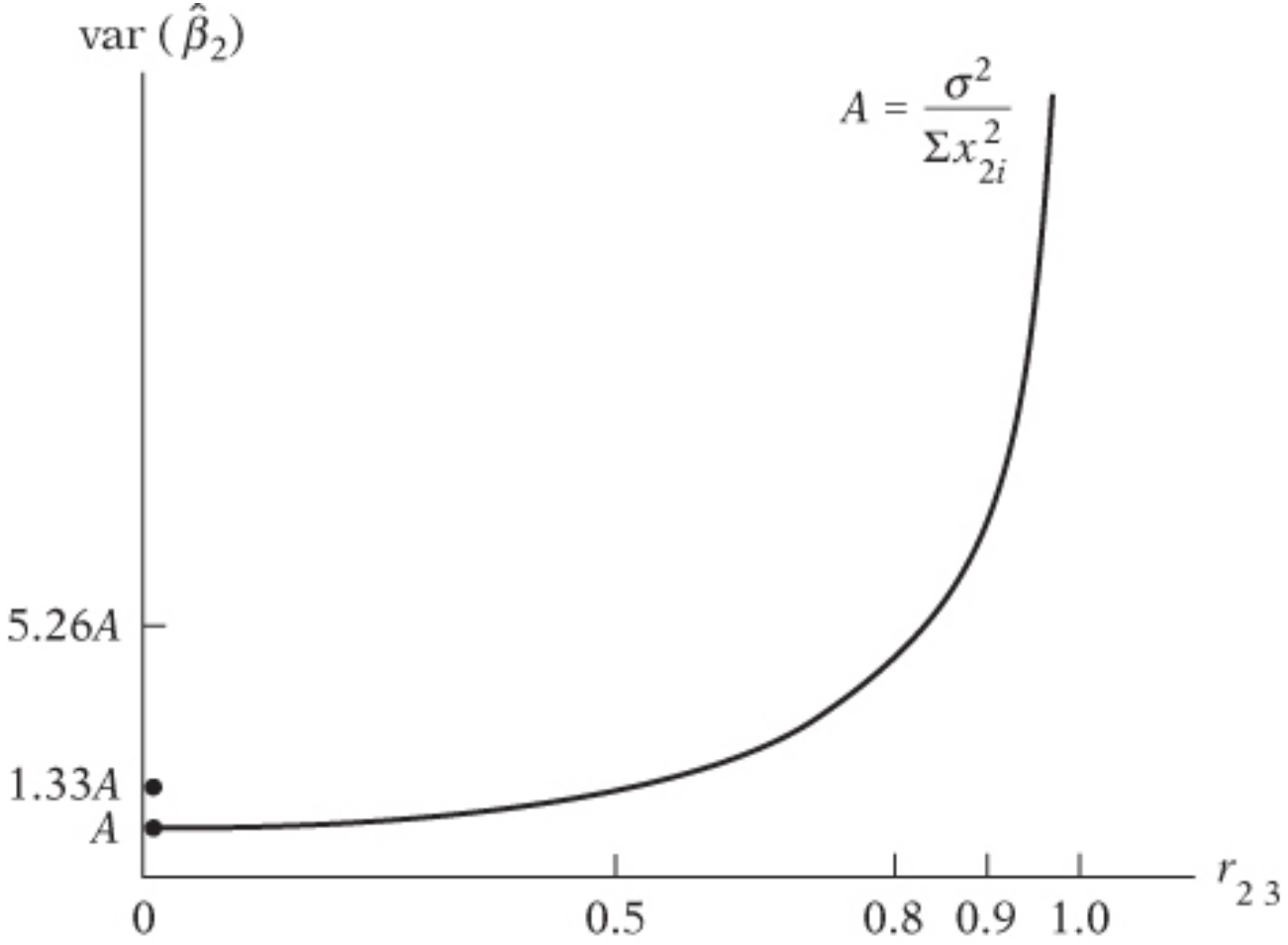
Value of r_{23}	VIF	var $(\hat{\beta}_2)$	$\frac{\text{var}\,(\hat{\beta}_2)(r_{23} \neq 0)}{\text{var}\,(\hat{\beta}_2)(r_{23} = 0)}$	cov $(\hat{\beta}_2, \hat{\beta}_3)$
(1)	(2)	(3)*	(4)	(5)
0.00	1.00	$\frac{\sigma^2}{\sum x_{2i}^2} = A$	—	0
0.50	1.33	$1.33 \times A$	1.33	$0.67 \times B$
0.70	1.96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

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

$\times$ = 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}

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k}$$

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

$\text{var}(\hat{\beta}_j)$ 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: $\hat{\beta}_2 \pm 1.96\ se\ (\hat{\beta}_2)$

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.

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

$$H_0 : \beta_i = 0$$

$$H_i : \beta_i \neq 0$$

$$CI : \hat{\beta}_i - se(\hat{\beta}_i) t_{\alpha/2} \leq \beta_i \leq \hat{\beta}_i + se(\hat{\beta}_i) t_{\alpha/2}$$

$r_{23} = 0$

$r_{23} \cdot 0.95$

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i}$

Reg X_{2i} X_{3i} X_{4i} X_{5i}

Tolerance (TOL)

- The inverse of the VIF is called tolerance (TOL)

$$TOL_j = \frac{1}{VIF_j} = (1 - R_j^2)$$

When $R_j^2 = 1$(perfect multicollinearity), $TOL_j = 0$.

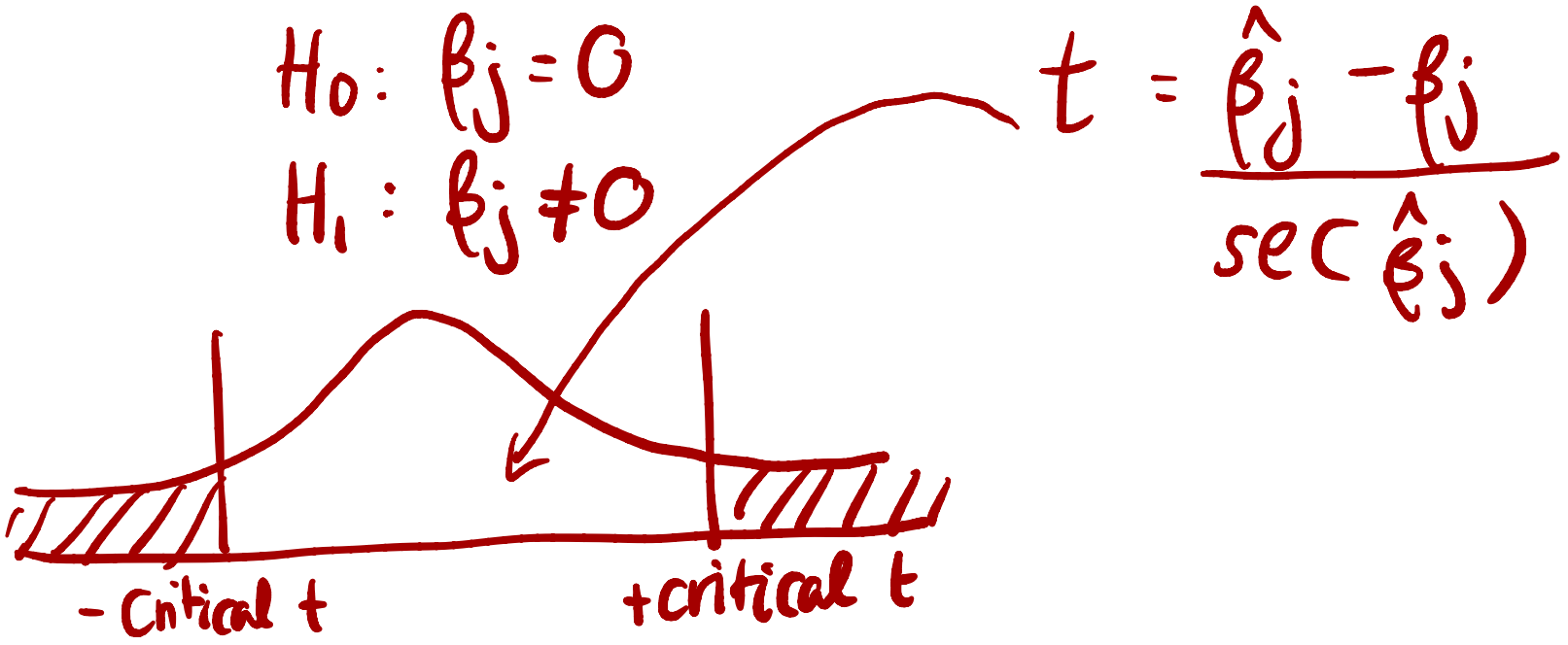
When $R_j^2 = 0$(no collinearity), $TOL_j = 1$.

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

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

- 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
- R-squared is high
- F test can reject the hypothesis that

H_1: otherwise

H_0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$

One of the signal of multicollinearity –insignificant t values but a high overall R-squared and a significant F value

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

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

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

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

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

$$R^2 = 0.8101$$

$$r_{23} = 0.5523$$

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

$$df = 5 - 3 = 2$$

Source	SS	df	MS
Model	8.10121951	2	4.05060976
Residual	1.89878049	2	.949390244
Total	10	4	2.5

Number of obs	5
F(2, 2)	4.27
Prob > F	0.1899
R-squared	0.8101
Adj R-squared	0.6202
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

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

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

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

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

$$R^2 = 0.8143$$

$$r_{23} = 0.8285$$

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

$$df = 5 - 3 = 2$$

Source	SS	df	MS
Model	8.14324324	2	4.07162162
Residual	1.85675676	2	.928378378
Total	10	4	2.5

Number of obs	5
F(2, 2)	4.39
Prob > F	0.1857
R-squared	0.8143
Adj R-squared	0.6286
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

Jan Kmenta, Elements of Econometrics, 2d ed., Macmillan, New York, 1986, p. 431

part II

Detection of Multicollinearity

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 an not of the population

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

Detection of Multicollinearity

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

$H_0: \beta_2 = \beta_3 = 0$

H_1: otherwise

$F_{cal} = 92.40$

$F_{crit}(2, 7) = ?$

Reject H_0.

∴ At least β will not equal to zero

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

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

$H_0: \beta_3 = 0$
$H_1: \beta_3 \neq 0$

$P_{val} > \alpha$ = 1%, 5%, 10%

Fail to reject H_0.

∴ $\beta_2 = 0$

$\beta_3 = 0$

Corr X_2 X_3 X_4 ....

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

- High zero order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even though the zero order or simple correlations are comparatively low

Example

Correlation between income (x2) and wealth (x3)

	x2	x3
x2	1.0000	
x3	0.9990	1.0000

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i}$$

Examination of partial correlations

A study of the partial correlations may be useful, there is no guarantee that they will provide an infallible guide to multicollinearity, for it may happen that both R^2 and all the partial correlations are sufficiently high

reg X_2 X_3 X_4 X_5

$R^2 = ?$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

Auxiliary regressions

$$F_i = \frac{R^2_{x_i x_2 x_3 \cdots x_k}/(k-2)}{(1-R^2_{x_i x_2 x_3 \cdots x_k})/(n-k+1)}$$

reg x_2 x_3 x_4 R^2

Follows the F distribution with k-2 and n-k+1 df.

n – number of sample

k – number of explanatory variables including the intercept term and

$R^2_{x_i x_2 x_3 \cdots x_k}$ is the coefficient of determination in the regression of variable X_i on the remaining X variables

$$X_{2i} = \gamma_2 + \gamma_3 X_{3i} + \gamma_4 X_{4i} + u_i \quad (R^2)$$

$H_0: \gamma_3 = \gamma_4 = 0$

H_1: otherwise

F-test

$$F = \frac{R^2 / (k-2)}{(1-R^2)/(n-k-1)}$$

Reject H_0

∴ The particular X_{2i} is collinear with other X_i'

- If the computed F exceeds the critical F_i at the chosen level of significance, it is taken to mean that the particular X_i is collinear with other X's

- If it does not exceed the critical F_i, we say that it is not collinear with other X's in which case we may retain that variable in the model

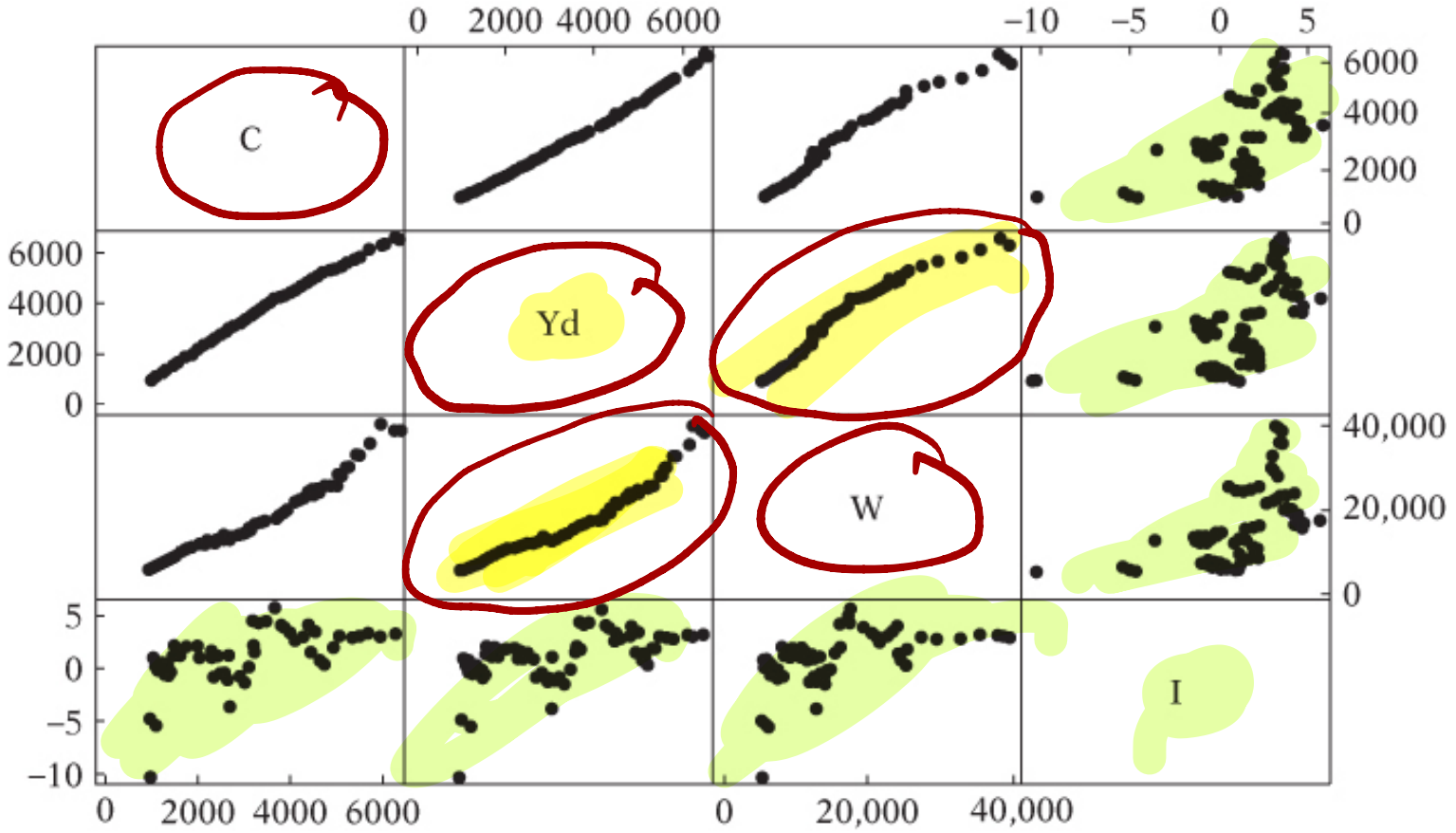
Tolerance and Variance Inflation Factor

- The larger the value of VIF_j, the more collinear the variable X_j **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 be highly collinear

- TOL_j a measure of multicollinearity in view of its intimate connection with VIF_j.
  - The closer TOL_j is to zero, the greater the degree of collinerity of that variable with the other regressors
  - The closer TOL_j is to 1, the greater the evidence that X_j is not collinear with the other regressors

C - Consumption
Yd - Real disposable income
W - Wealth
I - Real interest rate

Scatter plot



C – *Consumption*

Y_d – *Real disposable personal income*

W – *Real wealth*

I – *Real Interest Rate*

Remedial Measures

Remedial Measures

1. 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

A priori information

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

$$Y = consumption, X_2 = income, X_3 = wealth$$

$$\beta_3 = 0.10\beta_2$$

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

$$= \beta_1 + \beta_2 X_i + u_i$$

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

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, $\sum_i (X_{ji} - \bar{X}_j)^2$

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

Transformation of variables

- First difference form
- Ratio transformation

First difference form

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad (1)$$

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

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

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

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

Ratio transformation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad (1)$$

$$\frac{Y_t}{X_{3t}} = \beta_1 \left(\frac{1}{X_{3t}}\right) + \beta_2 \left(\frac{X_{2t}}{X_{3t}}\right) + \beta_3 + \left(\frac{u_t}{X_{3t}}\right) \qquad (2)$$

Ratio transformation, the error term $\left(\frac{u_t}{X_{3t}}\right)$

will be **heteroscedastic**, if the original error term is u_t 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

Examples

MULTICOLLINEARITY

Example 1: 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

$$Y_t = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$H_0: \beta_2 = \beta_3 = 0$ F-test
H_1 otherwise

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

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$
Fail to reject H_0.

$H_0: \beta_3 = 0$
$H_1: \beta_3 \neq 0$
Fail to reject H_0.

TOL

Variable	VIF	1/VIF
x2	**482.13**	**0.002074**
x3	**482.13**	**0.002074**
Mean VIF	**482.13**	

	x2	x3
x2	**1.0000**	
x3	**0.9990**	**1.0000**

$$\hat{Y}_i = 24.7747 + 0.9415X_{2i} - 0.0424X_{3i}$$

$$(6.7525) \quad (0.8229) \qquad (0.0807)$$

$$t = (3.6690) \ (1.1442) \qquad (-0.5261)$$

$$R^2 = 0.9635 \quad \bar{R}^2 = 0.9531 \qquad df = 10 - 3 = 7$$

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

F-test "Overall significance test"

$$H_0 = \beta_2 = \beta_3 = 0$$

H_1 = otherwise

$$F = \frac{4282.7770}{46.3494} = 92.4019$$

Reject the null hypothesis

(92.4019 > Critical F-value)

∴ At least one is not equal to zero.

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

r^2_{23} =

$$\hat{X}_{3i} = 7.5454 + 10.1909 X_{2i}$$

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

$$\hat{Y}_i = 24.4545 + 0.5091 X_{2i}$$

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

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

$$\hat{Y}_i = 24.411 + 0.0498 X_{3i}$$

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

$H_0: \beta_3 = 0$

$H_1: \beta_3 \neq 0$

Reject H_0.

Regressions show very clearly that in situations of extreme multicollinearity dropping the highly collinear variable will often make the other X variable statistically significant.

This result would suggest that a way out of extreme collinearity is to drop the collinearity variable.

Example 2: The Longley Data

The data are time series for the year 1947-1962 and pertain of

Y = number of people employed, in thousands

X_1 = GNP implicit price deflator

X_2 = GNP, millions of dollars

X_3 = number of people unemployed in thousands

X_4 = number of people in armed forces

X_5 = noninstitutionalized population over 14 years of age

X_6 = years, equal to 1 in 1947, 2 in 1948, and 16 in 1962

Source	SS	df	MS
Model	155088615	6	25848102.4
Residual	699138.24	8	87392.28
Total	155787753	14	11127696.6

Number of obs	15
F(6, 8)	295.77
Prob > F	0.0000
R-squared	0.9955
Adj R-squared	0.9921
Root MSE	295.62

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	-2.051082	8.70974	-0.24	0.820	-22.13578	18.03361
x2	-.0273342	.0331748	-0.82	0.434	-.1038355	.0491671
x3	-1.952293	.4767006	-4.10	0.003	-3.051567	-.8530199
x4	-.9582393	.2162271	-4.43	0.002	-1.45686	-.4596187
x5	.0513397	.233968	0.22	0.832	-.4881915	.5908709
x6	1585.156	482.6832	3.28	0.011	472.086	2698.225
_cons	67271.28	23237.42	2.89	0.020	13685.68	120856.9

	x1	x2	x3	x4	x5	x6
x1	1.0000					
x2	0.9937	1.0000				
x3	0.5917	0.5753	1.0000			
x4	0.4690	0.4588	-0.2033	1.0000		
x5	0.9833	0.9897	0.6748	0.3712	1.0000	
x6	0.9908	0.9948	0.6466	0.4222	0.9957	1.0000

Variable	VIF	1/VIF
x2	**1490.72**	**0.000671**
x6	**746.46**	**0.001340**
x5	**347.60**	**0.002877**
x1	**130.19**	**0.007681**
x3	**32.22**	**0.031034**
x4	**3.86**	**0.259149**
Mean VIF	**458.51**	

Source

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