ES1004 Econometrics by Example

Lecture 4: Multicollinearity

Dr. Hany Abdel-Latif

Swansea University, UK

Gujarati textbook, second edition

21st May 2016



Dr. Hany Abdel-Latif (2016)

ES1004ebe Lecture 3

Multicollinearity 1 / 23

CLRM Assumptions

- A₁: model is linear in parameters
- A₂: regressors are fixed non-stochastic
- **A**₃: the expected value of the error term is zero $E(u_i|X) = 0$
- **A**₄: homoscedastic or constant variance of errors $var(u_i|X) = \sigma^2$
- **A**₅: no autocorrelation, $cov(u_i, u_j) = 0, i \neq j$
- A_6 : no multicollinearity; no perfect linear relationships among the Xs
- A₇: no specification bias



Basic Idea

- CLRM assumes no exact linear relationship among explanatory variables A₆
- perfect multicollinearity
 - an exact relationship amongst the x's
 - is rarely encountered in practice, unless as a result of 'specification error' e.g., dummy variable trap
- imperfect multicollinearity
 - when explanatory variables are highly correlated
 - is a matter of degree
 - typically in macroeconomic time series data



Perfect Multicollinearity I

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

- if, for example, $X_{2i} + 3X_{3i} = 1$ we have perfect collinearity for $X_{2i} = 1 3X_{3i}$
- then we cannot include both X_{2i} and X_{3i} in the same regression model
- we cannot estimate the regression coefficients



Dr. Hany Abdel-Latif (2016)

Perfect Multicollinearity II

- examples of perfect collinearity
 - if we introduce income variables in both dollars and cents in the consumption function
 - dummy variable trap: when including as many dummies as the number of groups with the presence of the intercept
- in practice, exact linear relationships among regressors is a rarity



Imperfect Multicollinearity

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

- if we have $X_{2i} + 3X_{3i} + v_i = 1$ where v_i is a random term, for $X_{2i} = 1 3X_{3i} v_i$
- then we have imperfect multicollinearity
- no perfect linear relationship between the two variables
- in most cases, you we deal with imperfect (or near) collinearity rather than perfect collinearity



(日) (同) (三) (三)

Multicollinearity and OLS Estimation

- OLS estimators still BLUE
- high R^2 but will have insignificant coefficients
- regression coefficients are very sensitive to small changes in the data, especially of the sample is relatively small
- if two variables are highly collinear it is very difficult to isolate the impact of each variable separately on the regressand



Modelling Expenditure: Data

Expenditure (\$)	Income (\$)	Wealth (\$)
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	340	2435
150	260	2686



Dr. Hany Abdel-Latif (2016)

・ロト ・聞ト ・ヨト ・ヨト

Modelling Expenditure: Estimation

Dependent variable	Intercept	Income	Wealth	\mathbb{R}^2	
Expenditure	24.7747	0.9415	-0.0424	0.9635	
	(3.6690)	(1.1442)	(-0.5261)	Lingar	
Expenditure	24.4545	0.5091	Tr incontrol	0.9621	
and the second day of second law or	(3.8128)	(14.2432)	a Tanasan	Enghal .	
Expenditure	24.4410	Todachant	0.0498	0.9567	
al man of mus	(3.5510)	-	(13.2900)	1 martine	
Wealth	7.5454	10.1909	_	0.9979	
and the second second	(0.2560)	(62.0405)			



Dr. Hany Abdel-Latif (2016)

・ロト ・聞ト ・ヨト ・ヨト

Detection

Testing for Collinearity

- there is no unique test for multicollinearity
- **1** high R^2 but few significant t ratios
- high pairwise correlations among explanatory variables 2
- high partial coefficients 3
- significant F-test for auxiliary regressions
- bigh variance inflation factor [low tolerance factor]



Dr. Hany Abdel-Latif (2016)

Married Women's Hours of Work[.] Data

- Mroz (1987) Econometrica, 55, 765-99
- assessing the impact of several socio-economic variables
- data in Table 4.4 [see Piazza]
- cross-sectional data on 753 married women in 1975
- 325 married women did not work [i.e., zero hours of work]



Dr. Hany Abdel-Latif (2016)

Married Women's Hours of Work: Variables I

- hours 13 hours worked in 1975 [dependent variable]
- age 🕼 woman's age in years
- educ 🎼 years of schooling
- exper 🕼 actual labour market experience
- faminc 🆙 family income in 1975
- fathereduc 127 father's years of schooling
- hage 🎼 husband's age
- heduc 🕼 husband's years of schooling



Dr. Hany Abdel-Latif (2016)

ES1004ebe Lecture 3

Multicollinearity 12 / 23

Example 2

Married Women's Hours of Work. Variables II

- hhours I hours worked by husband
- hwage 12 husband's hourly wage, 1975
- kids618 🕼 number of kids between ages 6 and 18
- kidsl6 🕼 number of kids under age 6
- wage B estimated wage from earnings
- mothereduc 🕼 mother's years of education
- mtr 🕼 marginal tax rate facing a woman
- unemployment 12 unemployment rate in county of residence



Married Women's Hours of Work: A priori

- we would expect a
 - positive sign IP education, experience, father's education, mother's education
 - negative sign 1 age, husband's age, husband's hours of work, husband's wage, marginal tax rate, unemployment rate, number of kids under 6



Estimation

Dependent Variable: HOURS Method: Least Squares Date: 05/20/16 Time: 09:44 Sample: 1 753 IF HOURS>0 Included observations: 428

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C AGE EDUC EXPER FAMINC FATHEREDUC HAGE HEDUC HHOURS HWAGE KIDSL8 WAGE WAGE WAGE WATR	8595.360 -14.30741 -18.39847 -22.88057 0.013887 -7.471448 -5.566216 -6.769259 -0.473547 -141.7821 -24.50866 -191.5649 -48.14963 -1.837597 -6272.598 -6.772.598	1027.190 9.660582 19.34225 4.777417 0.006042 11.19227 8.938425 13.98780 0.073274 16.61801 28.06160 87.83197 10.41198 11.90008 10.85.438 10.63720	8.367843 -1.481009 -0.951206 4.789318 2.298543 -0.667554 -0.624966 -0.483940 -6.462701 -8.531837 -0.873388 -2.181038 -4.624447 -0.154419 -5.778864 -1.514984	0.0000 0.1394 0.3421 0.0000 0.5048 0.5323 0.6287 0.0000 0.3830 0.0297 0.0000 0.8774 0.0000 0.8774
R-squared Adjusted R-squared S.E. of regression Sum squared resid Log likelihood F-statistic Prob(F-statistic)	0.339159 0.315100 642.4347 1.70E+08 -3366.286 14.09655 0.000000	Mean depende S.D. depende Akaike info cr Schwarz crite Hannan-Quin Durbin-Watsc	lent var int var iterion rion n criter. in stat	1302.930 776.2744 15.80507 15.95682 15.86500 2.072493



э

Dr. Hany Abdel-Latif (2016)

Multicollinearity 15 / 23

イロト イヨト イヨト イヨト

Dependent Variable and Sample

Dependent Variable: HOURS Method: Least Squares Date: 05/20/16 Time: 09:44 Sample: 1 753 IF HOURS>0 Included observations: 428

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C AGE	8595.360	1027.190	8.367843	0.0000
EDUC	-18.39847	19.34225	-0.951206	0.1394
FAMINC	22.88057 0.013887	4.///41/ 0.006042	4.789318 2.298543	0.0000
FATHEREDUC	-7.471448	11.19227	-0.667554	0.5048
				ی الاقتصاد HECONO FSOCI

Insignificant Coefficients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
С	8595.360	1027.190	8.367843	0.0000
AGE	-14.30741	9.660582	-1.481009	0.1394
EDUC	-18.39847	19.34225	-0.951206	0.3421
EXPER	22.88057	4.777417	4.789318	0.0000
FAMINC	0.013887	0.006042	2.298543	0.0220
FATHEREDUC	-7.471448	11.19227	-0.667554	0.5048
HAGE	-5.586216	8.938425	-0.624966	0.5323
HEDUC	-6.769259	13.98780	-0.483940	0.6287
HHOURS	-0.473547	0.073274	-6.462701	0.0000
HWAGE	-141.7821	16.61801	-8.531837	0.0000
KIDS618	-24.50866	28.06160	-0.873388	0.3830
KIDSL6	-191.5649	87.83197	-2.181038	0.0297
WAGE	-48.14963	10.41198	-4.624447	0.0000
MOTHEREDUC	-1.837597	11.90008	-0.154419	0.8774
MTR	-6272.598	1085.438	-5.778864	0.0000
UNEMPLOYMENT	-16.11532	10.63729	-1.514984	0.1305
				FSOCIET
		•	D > < @ > < 문)	• ৰ≣ ► ≣ • গ ৭ (

Dr. Hany Abdel-Latif (2016)

Coefficient of Determination R^2

R-squared Adjusted R-squared S.E. of regression Sum squared resid	0.339159 0.315100 642.4347 1.70E+08	Mean dependent var S.D. dependent var Akaike info criterion Schwarz criterion	1302.930 776.2744 15.80507 15.95682
Log likelihood	-3366.286	Hannan-Quinn criter.	15.86500
F-statistic Prob(F-statistic)	14.09655 0.000000	Durbin-Watson stat	2.072493



Dr. Hany Abdel-Latif (2016)

ES1004ebe Lecture 3

Multicollinearity 18 / 23

イロト イヨト イヨト イヨト

Variance Inflation Factor VIF

Equation: UNTITLED Workfile: TABLE4_4::Table4_4									_ 🗆			
View	Proc	Object	Print	Name	Freeze		Estimate	Forecast	Stats	Resids		
	Representations											
	Estima	ation Out	:put									
	Actual	l,Fitted,F	Residua	I .	•							
	ARMA	Structur	re			L						
	Gradie	ents and	Derivat	ives	•	ŀ	Std. E	Error	t-Stati	istic	Prob.	
	Coyar	iance Ma	atrix				1027	.190	8.367	843	0.0000	
	Coefficient Diagnostics ▶					Scale	d Coefficie	ents				
	<u>R</u> esidu	ual Diagn	ostics		•		Confi	dence Inte	ervals			
	Stabili	ty Diagn	ostics		•		⊆onfi	dence Ellip	se			
	Label						⊻aria	nce Inflati	on Fact	ors		
					41 3 3 41	1	Coeff	icient Vari	ance <u>D</u>	ecompo:	sition	
	KIDS618 -24.5086 KIDS16 -191.564					ì	<u>W</u> ald	Test- Coe	fficient	Restrict	ions	
	1	WAGE		-41	3.14963		Omitt	ed Variabl	es Test	: - Likelih	ood Ratio	
MOTHEREDUC -1.837597 MTR -6272.598 UNEMPLOYMENT -16.11532							<u>R</u> edu	ndant Vari	iables T	'est - Lik	elihood Ra	atio
						2	Eacto	r Breakpo	int Tesl	t		

Dr. Hany Abdel-Latif (2016)

Multicollinearity 1

イロト イヨト イヨト イヨト

19 / 23

مجتمع الاقتصاد ECONOMIC SOCIETY

Variance Inflation Factor VIF

Variable	Coefficient	Uncentered	Centered
	Variance	VIF	VIF
C AGE EDUC EXPER FAMINC FATHEREDUC HAGE HEDUC HHOURS HWAGE KIDS618 KIDS618 KIDSL6 WAGE MOTHEREDUC MTR	1055118. 93.32684 374.1226 22.82372 3.65E-05 125.2668 79.89544 195.6586 0.005369 276.1581 787.4534 7714.456 108.4093 141.6118 1178175.	1094.176 176.2509 64.19296 5.555480 27.18584 12.10382 170.1046 34.13956 29.66169 18.59817 2.900083 1.383181 3.191149 14.90258 552.9496	NA 5.756163 2.021618 1.532452 5.144349 1.608908 5.224349 1.864803 1.887424 3.643849 1.410795 1.225962 1.229041 1.603344 7.215127



э

Dr. Hany Abdel-Latif (2016)

Multicollinearity

イロト イロト イヨト イヨト

20 / 23

How to Remedy for Collinearity

- what should we do when there is multicollinearity
 - nothing, for we often have no control over the data
 - redefine the model by excluding variables may attenuate the problem
 - cautious needed as to no omit relevant variables
- principal components analysis
 - construct artificial variables from regressors such that they are orthogonal to one another
 - these principal components becomes the regressors in the model
 - yet, the interpretation of the coefficients is not straightforward



Revised Women's Hours of Work

Dependent Variable: HOURS Method: Least Squares Date: 05/21/16 Time: 16:17 Sample: 1 753 IF HOURS>0 Included observations: 428

	Variable	Coefficient	Std. Error	t-Statistic	Prob.
-	C	8484.524	987.5952	8.591094	0.0000
	AGE	-17.72740	4.903114	-3.615540	0.0003
	EDUC	-27.03403	15.79456	-1.711604	0.0877
	EXPER	24.20345	4.653332	5.201315	0.0000
	FAMINC	0.013781	0.005866	2.349213	0.0193
	HHOURS	-0.486474	0.070462	-6.904046	0.0000
	HWAGE	-144.9734	15.88407	-9.126972	0.0000
	KIDSL6	-180.4415	86.36960	-2.089178	0.0373
	WAGE	-47.43286	10.30926	-4.600995	0.0000
	MTR	-6351.293	1029.837	-6.167278	0.0000
	UNEMPLOYMENT	-16.50367	10.55941	-1.562935	0.1188 🌽 🗢
					ر الاقتصاد WECON
					₽SOC
					er k ≣r i E i i





Dr. Hany Abdel-Latif (2016)

ES1004ebe Lecture 3

Multicollinearity 23 / 23