Multiple Regression Workshop #3-Problem of multicollinearity

ในเอกสารวิชาการนี้ เราจะมาทำความเข้าใจเกี่ยวกับกรณีที่ความสัมพันธ์ระหว่างตัวแปรอิสระด้วยกัน ทำให้แบบจำลองที่ใช้ตัวแปรอิสระในการพยากรณ์ค่าของตัวแปรตามอาจไม่มีประสิทธิภาพ

ตารางต่อไปนี้แสดงอัตราการเติบโตของยอดนำเข้า(import-ใช้สัญลักษณ์ Y) อัตราการเติบโตของผลิตภัณฑ์มวลรวมในประเทศ(Gross domestic product-GDP ใช้สัญลักษณ์ X1) และอัตราการเติบโตของเงินเฟ้อ (inflation-ใช้สัญลักษณ์ X2)

File Edit View Data Transform Analyze Graphs Utilities Add-ons Winc

1 : year 1

	year	Y	X1	X2	var
1	1	0.0540	0.0709	-0.1593	
2	2	0.0656	0.0505	-0.2683	
3	3	0.1475	0.0780	0.4801	
4	4	0.0686	0.0750	0.1348	
5	5	0.0455	0.0627	-0.0218	
6	6	0.0827	0.0464	0.1612	
7	7	-0.0157	0.0399	-0.2511	
8	8	0.0753	0.0640	-0.2611	
9	9	0.0841	0.0503	0.0527	
10	10	0.1540	0.0621	-0.1500	
11	11	0.0578	0.0432	0.0251	
12	12	0.0918	0.0600	-0.1119	
13	13	0.0949	0.0623	-0.0131	
14	14	0.0555	0.0585	-0.3613	
15	15	0.1593	0.0652	0.2579	
16					

ในที่นี้เราสนใจว่าการเติบโตในยอดนำเข้าจะขึ้นอยู่กับการเติบโตของผลิตภัณฑ์มวลรวมในประเทศ และอัตราเงินเฟ้ออย่างไร ดังนั้นรูปแบบสมการถดถอยพหุคูณที่เราต้องการคือ

$Y= a_O+ a_1 X1 + a_2 X2 + \varepsilon$

โดยที่ Y = growth of import

a_O = Y-intercept (ค่าคงที่)

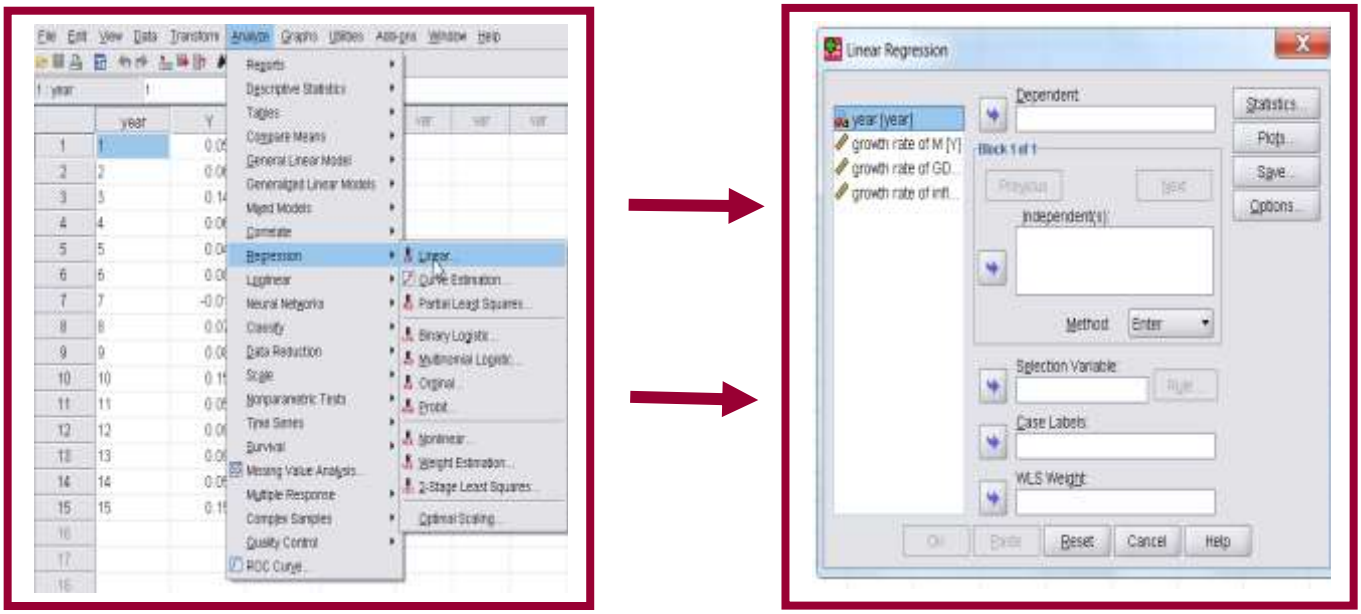
a_i = regression coefficients

X1 = อัตราการเติบโตของ GDP

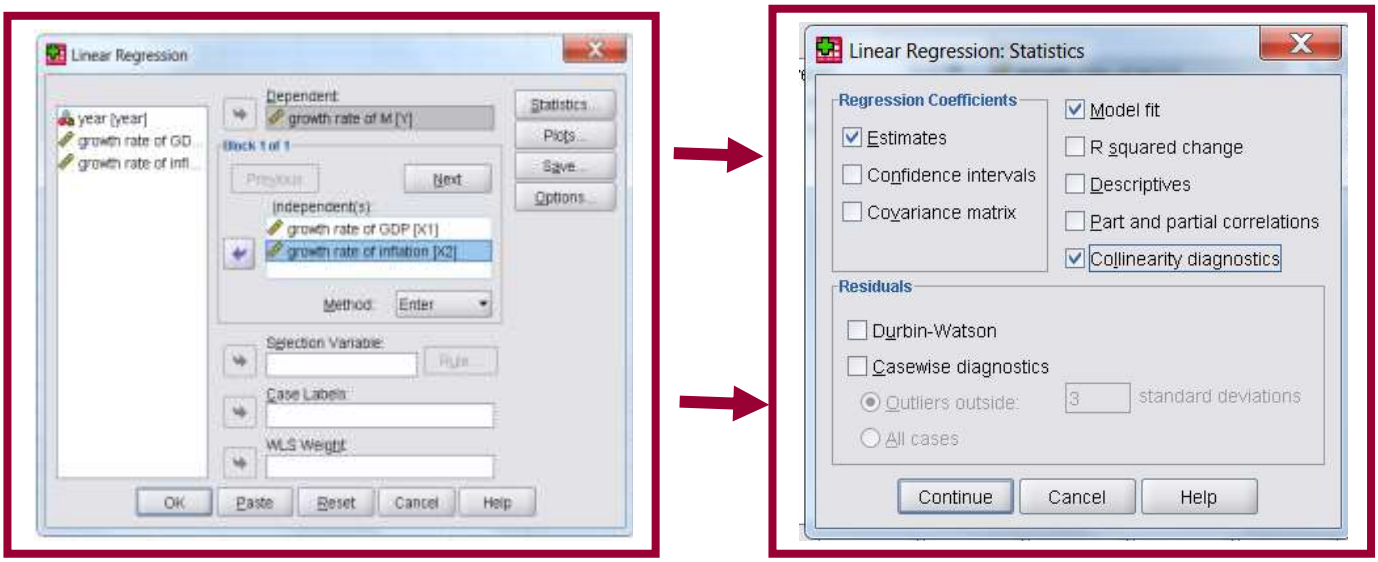
X2 = อัตราการเปลี่ยนแปลงในเงินเฟ้อ

ε = error terms

◉ไปที่ menu หลัก คลิ๊กAnalyze> Regression>Linear จะปรากฏ submenu ดังภาพด้านขวา



◉คลิ๊กเลือก Y ให้ไปอยู่ในช่อง Dependent X1 และ X2 ไปอยู่ในช่อง Independent(s) จากนั้นคลิ๊กปุ่ม Statistics จะปรากฏ submenu



◉คลิ๊กเลือก Estimates ภายใต้หัวข้อ Regression Coefficients และเลือก ช่อง Model fit และ Collinearity diagnostics

ภาพ 1 ต่อไปนี้แสดงcomputer output

ภาพที่ 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.650[a]	.423	.327	.0374421

a. Predictors: (Constant), growth rate of inflation, growth rate of GDP

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.012	2	.006	4.398	.037[a]
	Residual	.017	12	.001		
	Total	.029	14			

a. Predictors: (Constant), growth rate of inflation, growth rate of GDP

b. Dependent Variable: growth rate of M

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.002	.058		.029	.977		
	growth rate of GDP	1.393	.958	.344	1.453	.172	.859	1.164
	growth rate of inflation	.088	.047	.438	1.850	.089	.859	1.164

a. Dependent Variable: growth rate of M

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	growth rate of GDP	growth rate of inflation
1	1	2.008	1.000	.01	.01	.01
	2	.978	1.433	.00	.00	.84
	3	.014	11.990	.99	.99	.15

a. Dependent Variable: growth rate of M

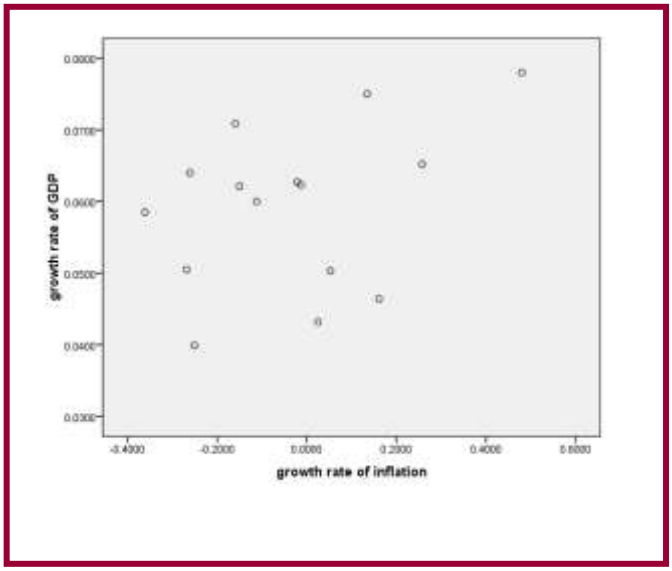
ผลการวิเคราะห์แสดงให้เห็นว่า ค่าของตัวแปรอิสระอันได้แก่ อัตราการเติบโตของ GDP และอัตราการเติบโตของเงินเฟ้อ ช่วยอธิบายความผันผวนในอัตราการเติบโตของสินค้าเข้าได้ร้อยละ 42.3 สมการถดถอยพหุคูณที่ได้มีนัยสำคัญทางสถิติร้อยละ 0.05 และสมการที่ได้คือ

Growth rate of M =0.002+ 1.393.growth rate of GDP +0.88.growth rate of inflation

(1.453) (1.850)

สิ่งที่น่าสังเกตคือ แม้ R-square จะบ่งบอกว่าตัวแปรอิสระใช้อธิบายความผันผวนของตัวแปรตามได้ในระดับปานกลาง และผลจากการทดสอบ F-test มีนัยสำคัญที่ α เท่ากับ 0.05 แต่กลับปรากฏว่าค่า regression coefficients ไม่มีนัยสำคัญทางสถิติที่ α =0.05 นี่คือลักษณะของตัวอย่างที่มีปัญหา multicollinearity

◉**พิจารณาความสัมพันธ์ระหว่างตัวแปรอิสระ**ด้วยกัน: ภาพด้านล่างเป็นกราฟแสดง scatter plot ระหว่างตัวแปรอิสระควบคู่กับข้อมูล Bivariate Correlation ระหว่างตัวแปรอิสระ



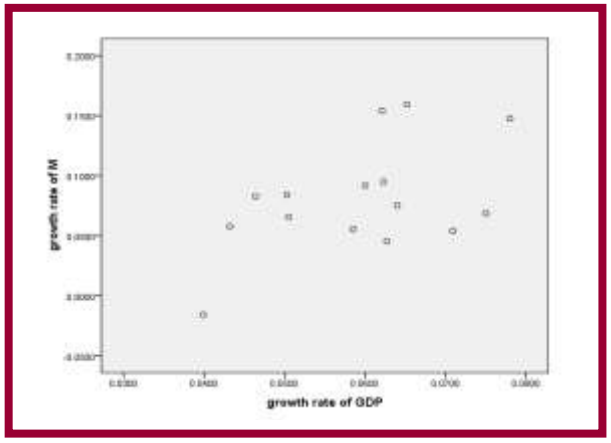
Correlations

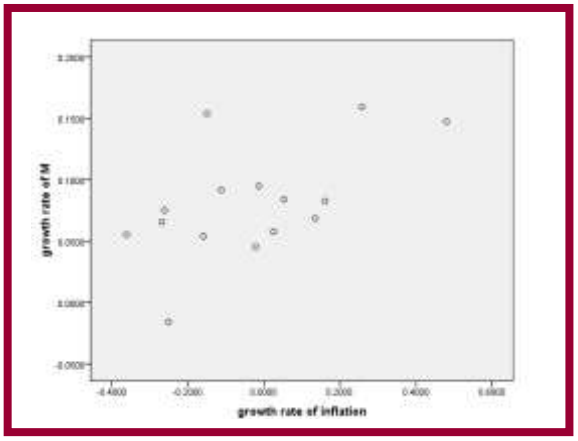
		growth rate of GDP	growth rate of inflation
growth rate of GDP	Pearson Correlation	1	.376
	Sig. (2-tailed)		.168
	N	15	15
growth rate of inflation	Pearson Correlation	.376	1
	Sig. (2-tailed)	.168	
	N	15	15

ภาพซ้ายมือแสดงให้เห็นกราฟแสดงความสัมพันธ์ระหว่างตัวแปรอิสระสองตัว ซึ่งจากภาพบ่งบอกว่าความสัมพันธ์เป็นเชิงบวก (positive relationship) แม้ความสัมพันธ์จะดูไม่เด่นชัด ส่วนภาพด้านขวาชี้ให้เห็น Pearson's Correlation ระหว่างตัวแปรอิสระทั้งสองซึ่งเป็นบวกแต่มีค่าอยู่ในระดับต่ำ (0.376) นี่เป็นข้อมูลที่ชี้ให้เห็นว่าเราอาจมีปัญหา multicollinearity อยู่จริง

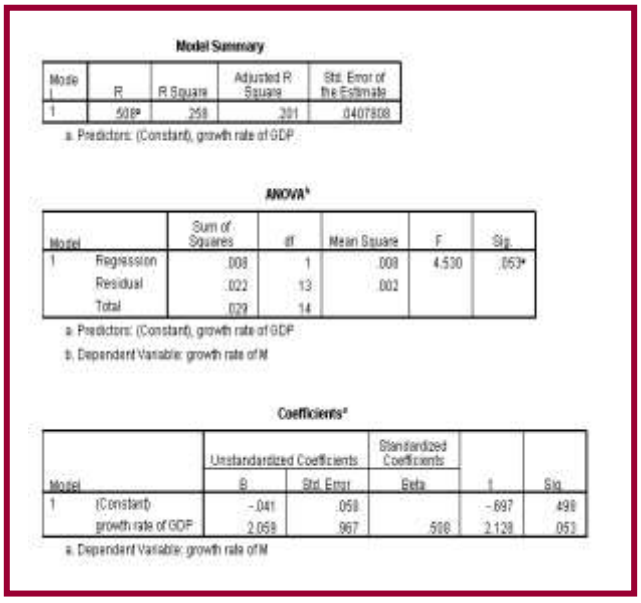
◉พิจารณาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระแต่ละตัว

ภาพด้านล่างซ้ายแสดง scatter plot หว่างตัวแปรอิสระตัวที่หนึ่งกับตัวแปรตาม ส่วนภาพขวาแสดง scatter plot ระหว่างตัวแปรอิสระตัวที่สองกับตัวแปรตาม





จาก scatter plot แสดงให้เห็นว่ามีความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระแต่ละตัว ภาพต่อไปแสดง computer output ของการถดถอยอย่างง่าย(simple regression) ที่มีระหว่างตัวแปรตามกับตัวแปรอิสระแต่ละตัว



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.508[a]	.258	.201	.0407808

a. Predictors: (Constant), growth rate of GDP

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.008	1	.008	4.530	.053[a]
	Residual	.022	13	.002		
	Total	.029	14			

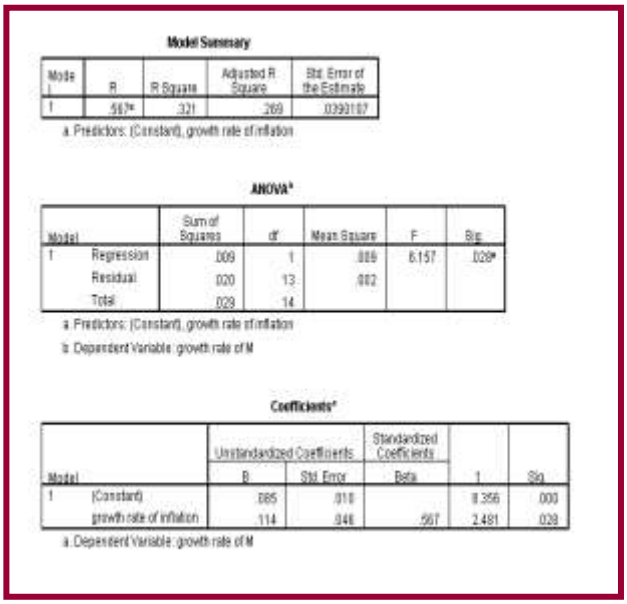
a. Predictors: (Constant), growth rate of GDP

b. Dependent Variable: growth rate of M

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.041	.059		-.697	.498
	growth rate of GDP	2.059	.967	.508	2.128	.053

a. Dependent Variable: growth rate of M



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.567[a]	.321	.269	.0393107

a. Predictors: (Constant), growth rate of inflation

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.009	1	.009	6.157	.028[a]
	Residual	.020	13	.002		
	Total	.029	14			

a. Predictors: (Constant), growth rate of inflation

b. Dependent Variable: growth rate of M

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.085	.010		8.356	.000
	growth rate of inflation	.114	.046	.567	2.481	.028

a. Dependent Variable: growth rate of M

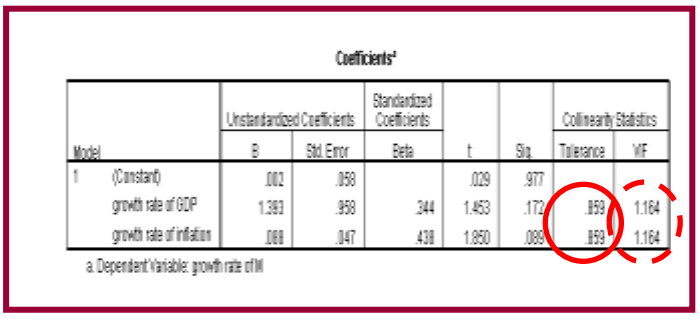
จาก computer output ในสองภาพข้างต้นเราพบว่า

1.Growth rate of GDP มีส่วนอธิบายความผันผวนของgrowth rate of M ร้อยละ 25.8 (R^2 =0.258) ในขณะที่อัตราการเปลี่ยนแปลงในราคาสินค้าหรือ growth rate of inflation มีส่วนอธิบายความผันผวนของgrowth Rate of M ร้อยละ 32.1(R^2=0.321)

2.การใช้ growth rate of GDP ในการพยากรณ์ growth rate of M ไม่มีนัยสำคัญทางสถิติแต่การใช้ growth rate of inflation ในการพยากรณ์ growth rate of M มีนัยสำคัญทางสถิติที่ระดับ α=0.05

3.เมื่อพิจารณาจากสมการถดถอยอย่างง่าย พบว่า regression coefficient ของ growth rate of GDP ไม่มีนัยสำคัญทางสถิติ แต่ regression coefficient ของ growth rate of inflation มีนัยสำคัญทางสถิติที่ระดับ α =0.05 กล่าวสั้นๆก็คือ หากให้เลือกเพียงตัวแปรอิสระเพียงตัวแปรเดียว เราจะเลือกใช้ growth rate of inflation ในการพยากรณ์ growth rate of M

◉ข้อมูล computer output ทั้งสิ้นมีส่วนสำคัญ ดังนั้นเราจะขอกลับไปที่ภาพที่ 2-A และภาพที่ 2-B ซึ่งเป็นส่วนที่ตัดทอน(truncated) จากภาพที่ 1 ด้านบน



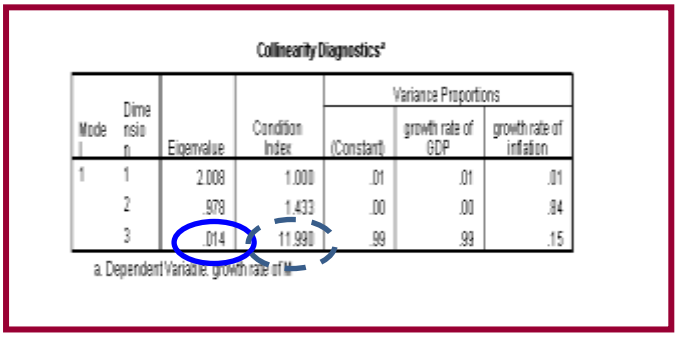
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.002	.958		.029	.977		
	growth rate of GDP	1.383	.958	.344	1.453	.172	.859	1.164
	growth rate of inflation	.088	.047	.438	1.850	.089	.859	1.164

a. Dependent Variable: growth rate of M

ภาพ 2-A

เมื่อพิจารณาจากค่า Tolerance ไม่มีค่าตัวแปรอิสระแม้แต่ตัวเดียวที่มีค่าต่ำกว่าค่า (1-R^2) หรือ0.577 ดังนั้น multicollinearity ไม่ใช่เรื่องที่น่ากังวล (Leech, Barrett และ Morgan,2005) เช่นเดียวกันเมื่อพิจารณาจากค่า VIF ตัวแปรอิสระทั้งสองตัวมีค่ามากกว่า 1 เล็กน้อย ดังนั้น multicollinearity อาจเป็นปัญหาบ้าง แต่ไม่ถึงขั้นต้องวิตก(Myers,1990 ; Bowerman & O'Connell,1990)



Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	growth rate of GDP	growth rate of inflation
1	1	2.008	1.000	.01	.01	.01
	2	.978	1.433	.00	.00	.84
	3	.014	11.990	.99	.99	.15

a. Dependent Variable: growth rate of M

จากตาราง Collinearity Diagnostics แสดงค่าของ Minimum Eigen value 0.014 ที่ต่ำมากซึ่งบ่งบอกว่าน่าจะมีปัญหา multicollinearity อยู่ ในส่วนของ Condition Index มีค่า 11.990 ดังนั้นปัญหาของ multicollinearity อยู่ในระดับต่ำ

ภาพ 2-B

คำศัพท์เฉพาะ

VIF:หรือ Variation Inflation Factor บ่งบอกว่า variance ของ coefficients ที่คำนวณได้เพิ่มพูนสูงขึ้นอันเป็นผลมาจากความสัมพันธ์ที่มีระหว่างตัวแปรอิสระหนึ่ง ๆกับตัวแปรอิสระตัวอื่น ๆ

Tolerance:มีค่าเท่ากับ 1/VIF

Eigen value: ใช้บ่งบอก variance ของ linear combination ของตัวแปร

Condition index: หรือ condition number of the correlation matrix ใช้วัด multicollinearity ของตัวแปรอิสระเป็นมวลรวม

Guidelines ที่กำหนดโดย guru

Belsley(1991), Shestra (2020)

- VIF หากมีค่าเท่ากับ 1 หมายความว่า ตัวแปรอิสระไม่มีความสัมพันธ์กัน
- หาก 1< VIF < 5 หมายความว่า ตัวแปรอิสระมีความสัมพันธ์กันในระดับปานกลาง
- หาก 5< VIF<10 หมายความว่า ตัวแปรอิสระมีความสัมพันธ์กันในระดับสูง
- หาก VIF>10 หมายความว่า regression coefficients หน้าตัวแปรอิสระไม่มีความน่าเชื่อถือเลย
- หาก eigen valueเข้าใกล้ 0.05 บ่งบอกว่าน่าจะมีปัญหา multicollinearity

Young (2017), Knoke (2002)

- CI(condition index) จะมีค่ามากกว่า 1 เสมอ
- CI< 15 หมายความว่า ตัวแปรอิสระมีความสัมพันธ์ในระดับต่ำ
- 15<CI<30 หมายความว่า ตัวแปรอิสระมีความสัมพันธ์ในระดับปานกลาง
- CI>30 หมายความว่า ตัวแปรอิสระมีความสัมพันธ์กันสูง

Bowerman & O'Connell (1990)

- ถ้า VIF โดยเฉลี่ยสูงกว่า 1 มีปัญหา multicollinearity
- Tolerance <0.1 multicollinearity อยู่ในระดับสูง

Menard (1995)

- Tolerance<0.2 multicollinearity น่าจะเป็นปัญหา

แนวทางแก้ปัญหาการมี multicollinearity

- ตัดตัวแปรอิสระที่มีความสัมพันธ์สูงออกไป
- สร้าง linear combination ของตัวแปรอิสระ
- ทำการหา principal component analysis เพื่อกลั่นกรององค์ประกอบที่ไม่มีความสัมพันธ์กัน จากนั้นใช้ partial least squares regression
- ใช้ Ridge regression ที่ช่วยให้สามารถจัดการปัญหา

Contribution this issue: ดร. ดนัย ปัตตพงศ์*

อยากเรียนรู้การนำสถิติข้างต้นนี้ไปใช้ในการวิจัยระดับสารนิพนธ์ (independent study) วิทยานิพนธ์ (thesis) ดุษฎีนิพนธ์(dissertation) ปรึกษาได้ที่ dpattaphongse@gmail.com

* ผู้แต่ง MBA's Made Easy (160+ issues) เอกสารวิชาการด้านศาสตร์การบริหารธุรกิจที่ช่วยให้ธุรกิจสามารถยืนหยัดและอยู่รอดได้ในภาวะที่โลกเปลี่ยนแปลงอยู่ตลอดเวลา

* ผู้พัฒนา FINALYSIS… a dedicated software สำหรับให้บริการนักธุรกิจที่ต้องการวิเคราะห์ความเป็นไปได้ทางการเงินของโครงการพัฒนาอสังหาริมทรัพย์ (บ้านจัดสรร/จัดสรรที่ดินเพื่อการอุตสาหกรรม/อาคารชุด/อาคารสำนักงานให้เช่า) โรงแรม โรงพยาบาลเอกชน ห้างสรรพสินค้า โรงงานน้ำตาล โรงงานกระดาษ โรงไฟฟ้าชีวมวล ฯลฯ ได้เห็นตัวเลขก่อนโครงการเกิด หลีกเลี่ยงความผิดพลาดเป็นร้อยเป็นพันล้านหากเกิดการลงทุนจริง(กำหนด DEBUT 1 เมษายน 2569)

* ผู้แต่งหนังสือ"การวิเคราะห์ความเป็นไปได้ทางการเงินและการจัดวงเงินเครดิตของโครงการลงทุน"ประกอบด้วยตัวอย่างของธุรกิจจริงที่ไม่เปิดเผยชื่อนับ 100 บริษัท ครอบคลุมอุตสาหกรรม 24 อุตสาหกรรม

*Co-developer ซอฟแวร์ en◉gex◉cel® สำหรับใช้ทดสอบ/เรียนรู้ศัพท์(ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 90 บทครอบคลุมศัพท์ระดับ SAT/IELTS/TOEFL กว่า 12,000 คำ) และไวยากรณ์อังกฤษ (ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 160 บทหรือกว่า 10,000 ข้อครอบคลุมเนื้อหาระดับอุดมศึกษาและTOEFL) มาพร้อมกับไฟล์เสียง/ไฟล์ข้อมูล/ฯลฯ อีกมาก(กำหนด DEBUT 1 เมษายน 2569)