

## Logistic Regression Workshop #1-Enter method

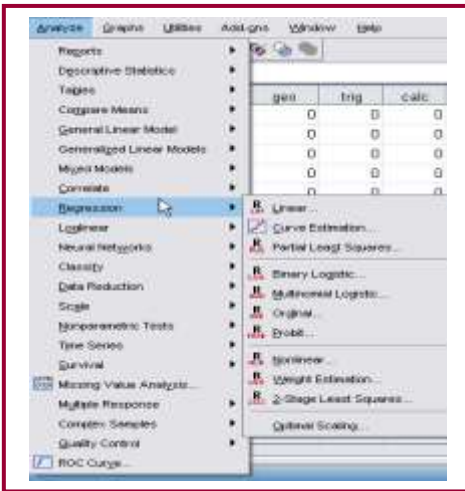
ภาพด้านล่างแสดงให้เห็นลักษณะของข้อมูลที่เราต้องการพยากรณ์การเลือกเรียนวิชาพีชคณิต2 โดยอาศัยข้อมูลจากเพศของนักเรียน ระดับการศึกษาของบิดา ผลคะแนนที่ได้จากการทดสอบวิชาว่าด้วยลวดลาย (mosaic test) และผลคะแนนที่ได้จากการทดสอบวิชาว่าด้วยการมองเห็น (visualization test)

	gender	faed	maed	alg1	alg2	geo	trig	calc	mathgr	grades	mathach	mosaic	visual	visual2	satm
1	1	10	10	0	0	0	0	0	0	4	9.00	31.0	8.75	7.00	500
2	1	2	2	0	0	0	0	0	0	5	10.33	56.0	4.75	4.00	460
3	1	2	2	0	0	0	0	0	1	6	7.67	25.0	4.75	3.00	420
4	0	3	3	1	0	0	0	0	0	3	5.00	22.0	1.00	1.00	400
5	1	3	3	0	0	0	0	0	0	3	-1.67	17.5	2.25	2.00	450
6	1	3	2	0	0	0	0	0	1	5	1.00	23.5	1.00	0.00	250
7	0	9	6	1	1	1	0	0	0	6	12.00	29.5	2.50	2.00	480
8	1	5	3	1	0	0	0	0	0	4	8.00	29.5	3.50	3.00	430
9	1	3	3	1	0	0	0	0	1	7	13.00	28.0	3.50	2.00	490
10	0	8	2	0	0	0	0	0	0	5	3.67	27.5	3.75	3.00	300
11	0	3	4	1	1	1	1	1	0	6	21.00	27.0	11.00	9.50	500
12	0	6	9	1	1	1	1	1	1	8	23.67	26.5	4.75	4.00	700
13	1	2	2	1	0	0	0	0	0	5	4.00	13.0	1.00	1.00	380
14	0	6	3	1	0	0	0	0	0	2	9.00	18.0	1.00	0.00	450
15	1	2	3	1	1	0	0	0	0	3	5.33	25.0	4.75	4.00	400
16	0	3	3	1	1	1	0	0	1	7	19.67	33.0	13.50	9.50	690
17	0	3	3	1	0	1	0	0	0	7	7.67	33.0	-0.25	1.00	430
18	0	9	7	1	0	1	0	0	0	5	14.33	27.0	6.00	5.00	500
19	0	3	3	1	1	1	0	0	1	6	14.33	26.0	9.75	9.50	510
20	1	8	3	1	1	1	1	0	1	8	19.67	25.0	8.75	9.50	500
21	0	5	5	1	0	1	0	0	0	5	17.00	41.0	-0.25	1.00	580
22	1	2	3	0	0	0	0	0	0	3	7.67	27.0	7.25	6.00	450
23	0	9	7	1	1	1	1	0	0	5	19.67	22.0	4.75	5.00	500
24	1	2	2	0	0	0	0	0	0	4	6.67	53.0	1.00	5.00	450
25	1	6	6	1	1	1	1	0	1	7	14.33	27.0	9.75	8.00	520
26	1	5	3	1	0	0	0	0	0	5	10.67	31.0	7.25	7.00	490

เนื่องจาก logistic regression มีความอ่อนไหวในประเด็นที่ตัวแปรอิสระหรือตัวแปรที่ใช้พยากรณ์ ( predictor variables) มี multicollinearity มาก ดังนั้นก่อนเริ่มการวิเคราะห์ด้วย logistic regression ผู้วิเคราะห์ต้องทำการตรวจสอบว่าตัวแปรอิสระมี multicollinearity ในระดับที่อาจเป็นปัญหาหรือไม่ก่อนเสมอ

เราจะใช้การวิเคราะห์การถดถอยพหุคูณ ( multiple regression) เพื่อตรวจสอบดูว่า จะมีปัญหา multicollinearity หรือไม่

- คลิกที่ Analyze>Regression>Linear



- คลิกเลือก algebra2 เป็นdependent variable และคลิกเลือกเพศ( gender)

การศึกษาของบิดา ( faed) ผลการทดสอบ mosaic test และผลการ

ทดสอบ visualization test เป็น independent variable



- คลิกที่ปุ่ม Statistics ในส่วน Regression

Coefficient เลือกCollinearity

diagnostics



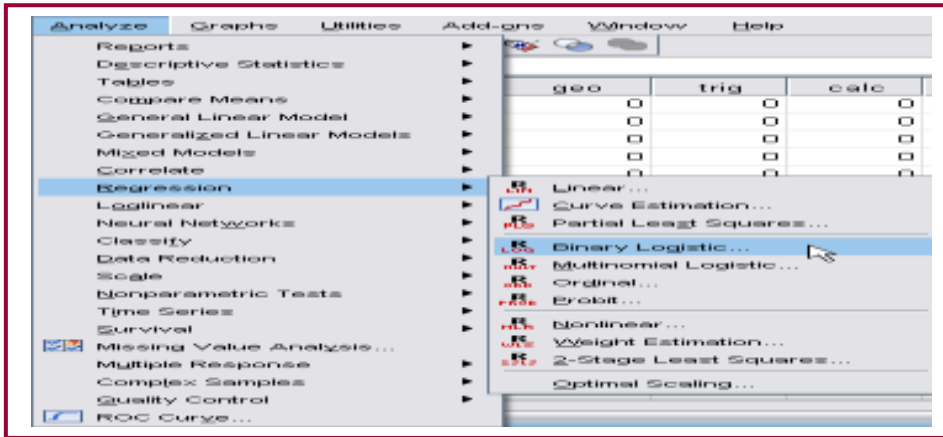
- คลิก Continue คลิก OK

ตารางด้านล่างแสดงให้เห็น computer output ที่เป็นส่วนหนึ่งของการวิเคราะห์การถดถอยพหุคูณ( multiple regression) จากค่าของ Tolerance และค่า VIF ไม่แสดงให้เห็นว่า multicollinearity จะเป็นปัญหาที่น่ากังวลแต่อย่างใด(ดูStatistics talks #8)

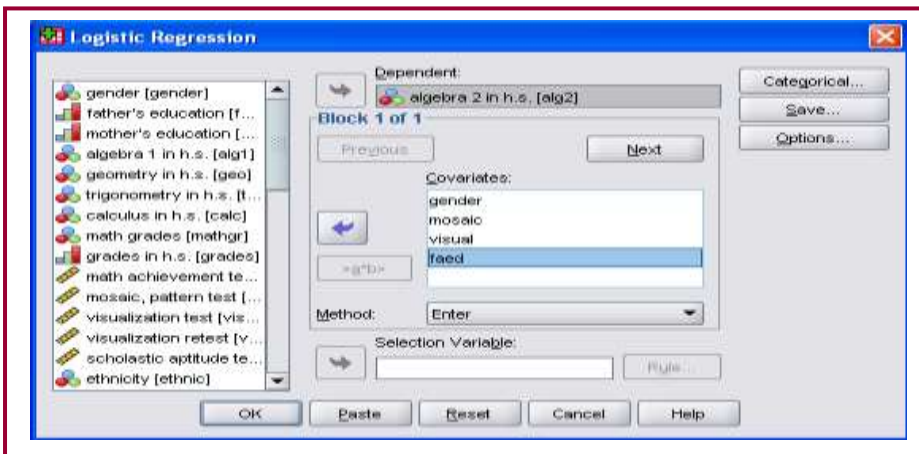
		Coefficients <sup>a</sup>					Collinearity Statistics	
		Unstandardized Coefficients	Standardized Coefficients	t	Sig.	Tolerance	VIF	
Model		B	Std. Error	Beta				
1	(Constant)	.204	.208		.981	.330		
	gender	-.089	.111	-.089	-.804	.424	.879	1.137
	father's education	.061	.019	.344	3.196	.002	.923	1.083
	mosaic, pattern test	-.006	.005	-.121	-1.162	.249	.994	1.006
	visualization test	.038	.014	.287	2.682	.009	.936	1.069

a. Dependent Variable: algebra 2 in h.s.

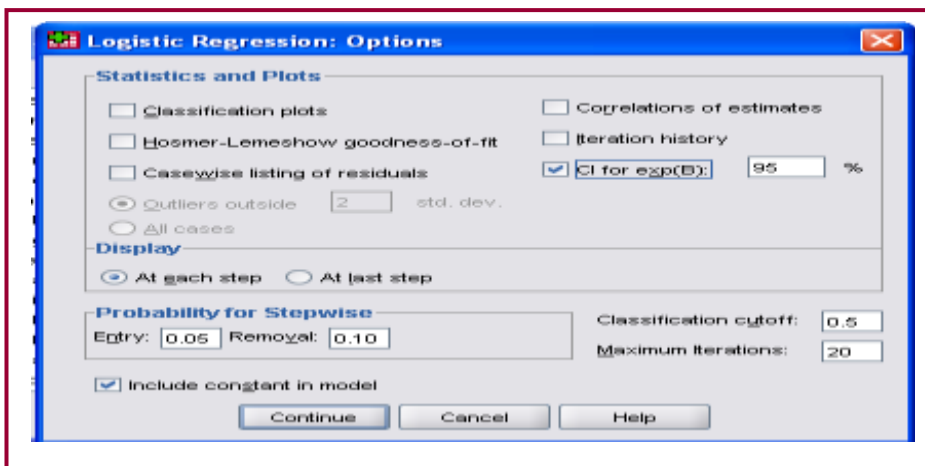
- คลิก Analyze>Regression>Binary Logistic



•คลิกเลือก algebra2 เป็น dependent variable คลิกเลือก gender mosaic visual และ faed ให้เข้าไปอยู่ในกล่อง Covariates ในworkshop นี้เราจะเลือกวิธี Enter คือใส่ตัวแปรอิสระเข้าไปพร้อมกันหมดในคราวเดียว ซึ่งขั้นตอนจะแตกต่างจากวิธีการstepwise ที่จะมีการทยอยนำตัวแปรอิสระเพิ่มเข้าไปทีละตัว



•คลิกที่ปุ่ม Options ในส่วนของ Statistics and Plots คลิกเลือก CI for exp(B) คลิก display at each step ในworkshop นี้เราจะเลือกวิธี Enter คือใส่ตัวแปรอิสระเข้าไปพร้อมกันหมดในคราวเดียว



คลิก Continue คลิก Ok

Computer output จะแบ่งออกเป็นส่วนๆดังนี้

**ส่วนที่หนึ่ง:** ข้อมูลเบื้องต้นแสดง Case Processing Summary และ Dependent Variable Encoding โดยในกรณีนี้ กำหนดให้เท่ากับ 1 ถ้ามีการเลือกวิชา algebra2 และเท่ากับ 0 หากไม่ได้เลือกวิชานี้

Unweighted Cases <sup>a</sup>		N	Percent
Selected Cases	Included in Analysis	73	97.3
	Missing Cases	2	2.7
	Total	75	100.0
Unselected Cases		0	.0
Total		75	100.0

a. If weight is in effect, see classification table for the total number of cases.

Origin...	Internal Value
not taken	0
taken	1

**ส่วนที่สอง:** แสดง Classification table โดยไม่อาศัยตัวแปรอิสระใด ๆ มาช่วยในการพยากรณ์( เป็น model ที่มีแต่ค่าคงที่เท่านั้น หรือเป็น baseline model) การพยากรณ์ใช้ข้อมูลสถิติเดิมคือ จากกลุ่มตัวอย่าง 73 คน ส่วนใหญ่ ( 39/73) ไม่ได้เลือกที่จะเรียน algebra 2 ส่วนตารางถัดไปแสดงข้อมูลของ model ที่เป็น baseline เท่านั้น

Observed		Predicted		Percentage Correct
		algebra 2 in h.s.		
		not taken	taken	
Step 0	algebra 2 in h.s.	39	0	100.0
	not taken	34	0	.0
Overall Percentage				53.4

สัดส่วนที่ถูกต้องโดยรวมคือ

$$(39+0) \times 100 / (39+34)$$

$$= 53.4$$

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	-.137	.235	.342	1	.559	.872

**ส่วนที่สาม:** เป็นตาราง Variables not in the Equation จะแสดงให้เห็นว่าหากมีการนำตัวแปรอิสระเข้ามา แต่ละตัวโดด ๆ จะมีส่วนช่วยในการอธิบายความแตกต่างระหว่างค่าที่สังเกตเห็น (observed) และค่าที่ได้จากการพยากรณ์ (predicted) มากน้อยเท่าใด ในที่นี้เพื่อ visualization test และระดับการศึกษาของบิดาล้วนมีความสัมพันธ์กับ algebra 2 และจะมีส่วนช่วยในการพยากรณ์

### Variables not in the Equation

	Score	df	Sig.
Step 0 Variables gender	4.765	1	.029
mosaic	.786	1	.375
visual	8.842	1	.003
faed	11.330	1	.001
Overall Statistics	19.771	4	.001

ส่วนที่สี่ : เป็นตาราง Omnibus Tests of Model Coefficients จากการนำเอาตัวแปรอิสระทุกตัวเข้ามาเป็นตัวแปรในการพยากรณ์ ( predictor variables) จะเห็นได้ว่า ค่าสถิติ  $\chi^2$  มีนัยสำคัญที่  $\alpha = 0.05$

### Block 1: Method =

#### Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step 1 Step	21.589	4	.000
Block	21.589	4	.000
Model	21.589	4	.000

ส่วนที่ห้า : แสดงให้เห็นข้อมูลของ -2LL ที่เกิดจาก model ที่รวมตัวแปรอิสระทุกตัวเข้าช่วยในการพยากรณ์ ซึ่งเท่ากับ 79.267 ลดลงจาก -2LL ของ model(baseline) ที่ใช้แต่ค่าคงที่ในการพยากรณ์ซึ่งเท่ากับ 100.857 (ไม่ได้ปรากฏให้เห็นใน computer output นี้) จำนวน 21.589 ซึ่งก็คือค่าสถิติ  $\chi^2$  ที่ปรากฏในตาราง Omnibus Tests นี้เอง ความหมายก็คือ การนำเอาตัวแปรอิสระเข้ามามีส่วนช่วยในการลดความคลาดเคลื่อนระหว่างข้อมูลที่ได้จากการสังเกตและข้อมูลที่ได้จากการพยากรณ์และมีนัยสำคัญทางสถิติ

### Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	79.267 <sup>a</sup>	.256	.342

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

อนึ่งจากค่าของ  $R^2_{CS}$  และ  $R^2_N$  ซึ่งให้เห็นคร่าวๆว่าตัวแปรอิสระสามารถอธิบายความผันผวนในค่าของตัวแปรตามหรือตัว

แปรผล (outcome variable) ได้พอสมควร

ส่วนที่หก : แสดง Classification Table จากการนำเอาตัวแปรอิสระทุกตัวเข้ามาพยากรณ์ พึ่งสังเกตว่าโดยรวมความสามารถในการพยากรณ์ได้ถูกต้องเพิ่มสูงขึ้น

Observed		Predicted		
		algebra 2 in h.s.		Percentage Correct
		not taken	taken	
Step 1	algebra 2 in h.s. not taken	31	8	79.5
	algebra 2 in h.s. taken	9	25	73.5
Overall Percentage				76.7

a. The cut value is .500

สัดส่วนความถูกต้องในการพยากรณ์สามารถคำนวณได้ดังนี้  
 $(31+25) \times 100 / (39+34) = 76.71$

ส่วนที่เจ็ด : แสดงให้เห็นผลจากการนำตัวแปรอิสระทุกตัวเข้ามาช่วยในการพยากรณ์ ในที่นี้เพศ (gender) และ mosaic ไม่ช่วยในการพยากรณ์ค่าของตัวแปรตามอย่างมีนัยสำคัญ แต่ก่อนหน้านี้ จากตารางที่แสดงในส่วนที่สามเราพบว่าเพศซึ่งเป็นตัวแปรอิสระมีความสัมพันธ์กับตัวแปรตามหากใช้โดด ๆ แสดงว่าเพศมีความสัมพันธ์กับตัวแปรอื่น ๆ ที่อยู่ใน model

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	gender	-.428	.575	.554	1	.457	.652
	mosaic	-.034	.031	1.207	1	.272	.967
	visual	.185	.077	5.773	1	.016	1.203
	faed	.290	.102	8.101	1	.004	1.337
	Constant	-1.335	1.106	1.457	1	.227	.263

a. Variable(s) entered on step 1: gender, mosaic, visual, faed.

อนึ่ง ตารางนี้แสดงให้เห็นว่า odd ของการเลือกเรียน algebra 2 เพิ่มขึ้นร้อยละ 33.7 หากเราทราบระดับการศึกษาของบิดาและ odd ของการเลือกเรียน algebra2 เพิ่มขึ้นร้อยละ 20.3 หากเราทราบผลของ visualization test

ผู้วิเคราะห์ที่เริ่มทำความคุ้นเคยกับ logistic regression อาจนึกว่า ในเมื่อพบว่าเพศและคะแนน mosaic test ไม่มีนัยสำคัญทางสถิติ จะตัดทิ้งไปเพื่อที่ผลการพยากรณ์จะได้มีความแม่นยำขึ้น คำตอบก็คือหากตัดทิ้งไป ความแม่นยำทางการพยากรณ์จะลดลง และค่า parameter ที่ได้มาจะต้องเปลี่ยนหมดไม่ใช่ค่าเดิม ดังนั้นตัวแปรอิสระบางตัวแม้ไม่มีนัยสำคัญ ก็ยังต้องคงเอาไว้

Contribution this issue: ดร. ดนัย บัณฑิตพงศ์

อยากเรียนรู้การนำสถิติข้างต้นนี้ไปใช้ในการวิจัยระดับสารนิพนธ์ (independent study) วิทยานิพนธ์ (thesis) ดุษฎีนิพนธ์ (dissertation) ปรึกษาได้ที่ [dpattaphongse@gmail.com](mailto: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 ซอฟต์แวร์ en@gex@cel<sup>®</sup> สำหรับใช้ทดสอบ/เรียนรู้ศัพท์(ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 90 บทครอบคลุมศัพท์ระดับ SAT/IELTS/TOEFL กว่า 12,000 คำ) และไวยากรณ์อังกฤษ (ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 160 บทหรือกว่า 10,000 ข้อครอบคลุมเนื้อหาในระดับอุดมศึกษาและTOEFL) มาพร้อมกับไฟล์เสียง/ไฟล์ข้อมูล/ฯลฯ อีกมาก(กำหนด DEBUT 1 เมษายน 2569)