

การวิเคราะห์การถดถอยจากการสำรวจตัวอย่างที่มีการปรับ ตามแบบแผนการชักตัวอย่างด้วย PROC SURVEYREG Regression Analysis from Sample Survey Data with respect to the Sample Designs by PROC SURVEYREG

ภคณัฐ ประกอบใส และ กมล นุชบา

ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี

มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ปทุมธานี 12121

1. คำนำ

โดยทั่วไปแล้วนักวิจัยทำการศึกษาลักษณะที่สนใจของประชากรโดยการสำรวจจากตัวอย่าง และเลือกใช้แผนแบบการชักตัวอย่างที่เหมาะสมกับลักษณะของประชากร เช่น ใช้แผนแบบการชักตัวอย่างแบบมีชั้นภูมิ (stratification) หรือใช้แผนแบบการชักตัวอย่างที่กำหนดน้ำหนักในการชักตัวอย่างที่ต่างกัน เป็นต้น เพื่อที่จะทำการอ้างอิงเกี่ยวกับประชากรได้อย่างถูกต้องตรงตามความเป็นจริง ซึ่งการทำการอ้างอิงที่ถูกต้องในเชิงสถิติ นั้น นักวิจัยจะต้องทำการวิเคราะห์ข้อมูลให้สอดคล้องกับแผนแบบการชักตัวอย่างที่ใช้ด้วย [1]

การวิเคราะห์การถดถอยเป็นวิธีการหนึ่งที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณ โดยมีจุดประสงค์เพื่ออธิบายหรือพยากรณ์ตัวแปรตัวหนึ่ง ด้วยตัวแปรตัวหนึ่งหรือมากกว่าหนึ่งตัว ซึ่งส่วนใหญ่แล้วนักวิจัยจะใช้โปรแกรม SPSS โดยใช้เมนู Analyze ตามด้วยเมนู Regression หรือใช้โปรแกรม SAS โดยใช้ PROC GLM หรือ PROC REG เป็นต้น อย่างไรก็ตามคำสั่งเหล่านี้จะทำการคำนวณสถิติต่างๆ ภายใต้ข้อสมมติว่าตัวอย่างนั้นเป็นตัวอย่างสุ่มแบบง่าย (simple random sample) จากประชากรที่มีขนาดอนันต์ (infinite population) ซึ่งโดยทั่วไป คำสั่งเหล่านี้ให้ค่าประมาณความแปรปรวนของตัวประมาณที่ไม่ถูกต้องหากนำไปใช้กับตัวอย่างที่ได้มา

จากแผนแบบการชักตัวอย่างอื่นๆ ที่ไม่ใช่แผนแบบการชักตัวอย่างสุ่มแบบง่าย [1 และ 2] ซึ่งปัญหาดังกล่าวในปัจจุบันสามารถแก้ไขได้โดยใช้คำสั่งที่มีการพัฒนาขึ้นมาใหม่ในโปรแกรมสำเร็จรูป SAS ตั้งแต่รุ่น 7.0 ขึ้นไป (รุ่นล่าสุดคือรุ่น 9.1.3)

2. PROC SURVEYREG

PROC SURVEYREG ใน SAS ใช้สำหรับวิเคราะห์การถดถอยของข้อมูลที่ได้จากการสำรวจจากตัวอย่าง โดยมีการปรับให้สอดคล้องกับแบบแผนการชักตัวอย่าง โปรแกรมจะทำการกำหนดตัวแบบเชิงเส้นให้กับข้อมูล จากนั้นจะทำการประมาณค่าสัมประสิทธิ์การถดถอย ทำการทดสอบสมมติฐานต่างๆ และพยากรณ์ค่าที่สนใจของข้อมูลที่ได้จากการสำรวจ

การประมาณค่าสัมประสิทธิ์การถดถอยใน PROC SURVEYREG ใช้วิธีกำลังสองน้อยที่สุดวางนัยทั่วไป (generalized least-squares) โดยถ้าประชากรมีการแบ่งเป็นกลุ่มหรือมีการใช้หน่วยตัวอย่างขั้นต้น (primary sampling units) ในการเลือกตัวอย่างจะสมมติว่าค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากันในทุกชั้นภูมิหรือทุกหน่วยตัวอย่างขั้นต้น ในการประมาณเมตริกซ์ความแปรปรวน-ความแปรปรวนร่วมของสัมประสิทธิ์การถดถอย PROC SURVEYREG จะใช้วิธีการกระจาย

อนุกรมเทเลอร์ (Taylor expansion) เพื่อประมาณความคลาดเคลื่อนจากตัวอย่าง (sampling errors) ของตัวประมาณที่สอดคล้องกับแบบแผนการเลือกตัวอย่างนั้นๆ วิธีนี้จะทำให้ได้ค่าประมาณเชิงเส้น (linear approximation) ของตัวประมาณ จากนั้นจะใช้ตัวประมาณความแปรปรวนสำหรับการประมาณค่าเชิงเส้นนี้เพื่อประมาณค่าความแปรปรวนของตัวประมาณอีกทีหนึ่ง [4]

2.1 การชักตัวอย่างสุ่มแบบง่าย

ตัวอย่างต่อไปนี้เป็นการใช้ PROC SURVEYREG เพื่อศึกษาความสัมพันธ์ระหว่างผลผลิตข้าวโพดและขนาดของพื้นที่เพาะปลูกโดยมีประชากรคือไร่ข้าวโพดจำนวน 235 แห่งในรัฐ Iowa และรัฐ Nebraska ถ้าทำการชักตัวอย่างไร่ข้าวโพดจำนวน 19 แห่ง โดยใช้วิธีการชักตัวอย่างสุ่มแบบง่าย (SRS) จากนั้นทำเก็บข้อมูลพื้นที่ในการเพาะปลูกข้าวโพด (ตัวแปร FarmArea หน่วยเป็นไร่) และผลผลิตข้าวโพดต่อปี (ตัวแปร ComYield หน่วยเป็นตัน) จากไร่ที่ตกเป็นตัวอย่าง โปรแกรมที่ใช้ในการวิเคราะห์การถดถอยและพยากรณ์ผลผลิตข้าวโพดทั้งหมด ถ้าทราบว่พื้นที่เพาะปลูกทั้งหมดคือ 21,950 ไร่ แสดงในภาพที่ 1 ซึ่งการพยากรณ์หรือประมาณยอดรวมในที่นี้ใช้ตัวประมาณแบบถดถอย (regression estimator) โดยมีจำนวนไร่เป็นตัวแปรช่วยในการประมาณ

จากภาพที่ 1 คำสั่ง TOTAL= 235 ใน PROC SURVEYREG เป็นการระบุจำนวนประชากรทั้งหมดซึ่งโปรแกรมจะใช้นี้ค่านี้เพื่อทำการปรับค่าประมาณความแปรปรวนอันเนื่องมาจากการชักตัวอย่างจากประชากรที่มีขนาดจำกัด ส่วนคำสั่ง MODEL เป็นการระบุตัวแบบเชิงเส้นที่ใช้ คำสั่งย่อย CLPARM เป็นการสั่งให้โปรแกรมคำนวณช่วงความเชื่อมั่นของพารามิเตอร์ ส่วนคำสั่ง ESTIMATE ในที่นี้จะพยากรณ์ผลผลิตข้าวโพดของพื้นที่เพาะปลูกทั้งหมด คำสั่งนี้จะต้องมีการสร้างป้ายชื่อ (label)

ทุกครั้ง ซึ่งในที่นี้คือ 'ComYield' ส่วน INTERCEPT 235 เป็นการระบุตัวแปรช่วยในการประมาณยอดรวมตามแนววิธีของตัวประมาณแบบถดถอยซึ่งเท่ากับ 235 ไร่ ส่วน FarmArea 21950 เป็นการระบุพื้นที่เพาะปลูกของไร่ข้าวโพดที่ต้องการพยากรณ์ผลผลิตข้าวโพด ผลลัพธ์ที่ได้แสดงในภาพที่ 2

จากภาพที่ 2 ขนาดของพื้นที่เพาะปลูกมีผลต่อผลผลิตข้าวโพดจริง (p-value ในตาราง ANOVA เท่ากับ 0.0042) ส่วนการทดสอบนัยสำคัญของเทอมต่างๆ ในตัวแบบพบว่าตัวแปร FarmArea มีนัยสำคัญ (p-value เท่ากับ 0.0154) ผลลัพธ์ในภาพที่ 3 ค่าประมาณสัมประสิทธิ์การถดถอยมีค่า 0.184 โดยมีค่าคลาดเคลื่อนมาตรฐานเท่ากับ 0.069 ตัน สรุปได้ว่าถ้าเพิ่มพื้นที่

```
DATA Farms;
  INPUT State $ Region FarmArea ComYield @@;
  DATALINES;
Iowa 1 100 54 Iowa 1 83 25 Iowa 1 25 10
Iowa 2 120 83 Iowa 2 50 35 Iowa 2 110 65
Iowa 2 60 35 Iowa 2 45 20 Iowa 3 23 5
Iowa 3 10 8 Iowa 3 350 125 Nebraska 1 130 20
Nebraska 1 245 25 Nebraska 1 150 33 Nebraska 1 263 50
Nebraska 1 320 47 Nebraska 1 204 25 Nebraska 2 80 11
Nebraska 2 48 8
;
PROC SURVEYREG DATA=Farms TOTAL=235;
  MODEL ComYield = FarmArea / CLPARM ;
  ESTIMATE 'ComYield' INTERCEPT 235 FarmArea 21950;
RUN;
```

ภาพที่ 1 คำสั่ง SAS ที่ใช้วิเคราะห์การถดถอยสำหรับแผนแบบการชักตัวอย่างสุ่มแบบง่ายจากประชากรขนาดจำกัด

ANOVA for Dependent Variable CornYield					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6388.90	6388.903	10.93	0.0042
Error	17	9939.10	584.653		
Corrected Total	18	16328.00			
Tests of Model Effects					
Effect	Num DF	F Value	Pr > F		
Model	1	7.16	0.0154		
Intercept	1	2.98	0.1017		
FarmArea	1	7.16	0.0154		

ภาพที่ 2 ตารางวิเคราะห์ความแปรปรวนและการทดสอบอิทธิพลของทอมต่างๆในตัวอย่าง (จากคำสั่งในภาพที่ 1)

ในการเพาะปลูก 1 ไร่ จะทำให้ได้ผลผลิตเพิ่มขึ้นโดยเฉลี่ย 0.184 ตันหรืออยู่ในช่วง 0.040 ตัน ถึง 0.328 ตัน ด้วยความเชื่อมั่นร้อยละ 95 ส่วนค่าประมาณของผลผลิตข้าวโพดเมื่อทราบพื้นที่ในการเพาะปลูกทั้งหมดเป็น 21,950 ไร่โดยวิธีตัวประมาณแบบถดถอยคือ 7003.04 ตัน หรืออยู่ในช่วง 4,785.26 ตัน ถึง 9,220.82 ตัน ด้วยความเชื่อมั่นร้อยละ 95 โดยมีค่าคลาดเคลื่อนมาตรฐานของการประมาณค่าเท่ากับ 1,055.62 ตัน

Estimated Regression Coefficients					
Parameter	Estimate	Standard Error	t Value	Pr > t	95% Confidence Interval
Intercept	12.64378	7.3296	1.73	0.1017	-2.7552 28.0427
FarmArea	0.18368	0.0686	2.68	0.0154	0.0395 0.3279
Analysis of Estimable Functions					
Parameter	Estimate	Standard Error	t Value	Pr > t	95% Confidence Interval
CornYield	7003.0397	1055.6229	6.63	<.0001	4785.26 9220.82

ภาพที่ 3 สัมประสิทธิ์การถดถอยและค่าพยากรณ์ผลผลิตข้าวโพด (จากคำสั่งในภาพที่ 1)

2.2 การชักตัวอย่างสุ่มแบบมีชั้นภูมิ

ถ้าสมมติว่าไร่ข้าวโพดจำนวน 19 แห่งในตัวอย่างก่อนหน้านี้จริงๆ แล้วเป็นตัวอย่างที่ได้จากการชักตัวอย่างสุ่มแบบมีชั้นภูมิกล่าวคือประชากรไร่ข้าวโพดทั้งหมด 235 แห่ง แบ่งเป็นชั้นภูมิตามรัฐและเขตดังตารางที่ 1 จากนั้นทำการชักตัวอย่างสุ่มแบบง่ายจากแต่ละชั้นภูมิ และได้ตัวอย่างรวมทั้งหมด 19 แห่ง

ตารางที่ 1 จำนวนไร่ข้าวโพดจำแนกตามรัฐและเขต

ชั้นภูมิ	รัฐ	เขต	จำนวนไร่ใน		เนื้อที่เพาะปลูกรวม
			ประชากร	ตัวอย่าง	
1	IA	1	100	3	13,200
2	IA	2	50	5	
3	IA	3	15	3	
4	NE	1	30	6	8,750
5	NE	2	40	2	
รวม			235	19	21,950

ในกรณีที่อัตราการชักตัวอย่างในแต่ละชั้นภูมิไม่เท่ากัน จำเป็นต้องมีการคำนวณน้ำหนักของการชักตัวอย่าง (sampling weight) เพื่อใช้ในการคำนวณค่าสถิติที่ไม่เอนเอียง ในตัวอย่างนี้น้ำหนักของการชักตัวอย่างได้แก่ ส่วนกลับ (reciprocals) ของค่าความน่าจะเป็นของการเลือกตัวอย่าง ดังนั้นจึงต้องสร้างตัวแปร Weight ซึ่งแทนน้ำหนักของการชักตัวอย่างดังกล่าวในภาพที่ 4

DATA Farms;

INPUT State \$ Region FarmArea CornYield Weight ๑๑;

DATALINES;

Iowa 1 100 54 33.333	Iowa 1 83 25 33.333
Iowa 1 25 10 33.333	Iowa 2 120 83 10.000
Iowa 2 50 35 10.000	Iowa 2 110 65 10.000
Iowa 2 60 35 10.000	Iowa 2 45 20 10.000
Iowa 3 23 5 5.000	Iowa 3 10 8 5.000
Iowa 3 350 125 5.000	Nebraska 1 130 20 5.000
Nebraska 1 245 25 5.000	Nebraska 1 150 33 5.000
Nebraska 1 263 50 5.000	Nebraska 1 320 47 5.000
Nebraska 1 204 25 5.000	Nebraska 2 80 11 20.000
Nebraska 2 48 8 20.000	

ภาพที่ 4 คำสั่ง SAS ที่ใช้อ่านชุดข้อมูลซึ่งระบุค่าน้ำหนักของการชักตัวอย่างในแต่ละชั้นภูมิ

จากภาพที่ 4 น้ำหนักในการเลือกตัวอย่าง ในรัฐ Iowa เขต 1 คือ 33.333 ซึ่งคำนวณจากส่วนกลับของค่าความน่าจะเป็นของการชักตัวอย่าง โดยที่ความน่าจะเป็นของการชักตัวอย่างเท่ากับขนาดตัวอย่างหารด้วยขนาดประชากรในแต่ละชั้นภูมิ

การวิเคราะห์ข้อมูลจากตัวอย่างแบบมีชั้นภูมิจากประชากรที่มีขนาดจำกัดด้วยโปรแกรม SAS จะต้องระบุขนาดของประชากรในแต่ละชั้นภูมิใน PROC SURVEYREG ในที่นี้จะใช้ชุดข้อมูลชื่อ StratumTotals ในการระบุขนาดของชั้นภูมิ ดังคำสั่งในภาพที่ 5

จากภาพที่ 5 ตัวแปร _TOTAL_ เป็นตัวแปรที่ใช้แทนจำนวนไร้ข้าวโพดทั้งหมดในแต่ละรัฐซึ่งต้องตั้งชื่อตัวแปรโดยใช้คำว่า _TOTAL_ เท่านั้น จากนั้นโปรแกรมจะใช้ขนาดของชั้นภูมิในการปรับค่าประมาณความแปรปรวนอันเนื่องมาจากการชักตัวอย่างจากประชากรที่มีขนาดจำกัด ถ้าผู้ใช้ไม่ระบุขนาดของประชากรหรืออัตราการชักตัวอย่าง (sampling rate) จะไม่มีการปรับแก้ไขเนื่องจากประชากรที่มีขนาดจำกัด (finite population correction) ในการคำนวณสถิติต่างๆ จากนั้นให้ทำการ

วิเคราะห์ข้อมูลจากการชักตัวอย่างแบบมีชั้นภูมิตามคำสั่งในภาพที่ 6

DATA StratumTotals;

INPUT State \$ Region _TOTAL_ ๑๑;

DATALINES;

Iowa 1 100	Iowa 2 50	Iowa 3 15
Nebraska 1 30	Nebraska 2 40	

ภาพที่ 5 คำสั่ง SAS ที่ใช้สร้างชุดข้อมูลที่ระบุค่าของชั้นภูมิและ ขนาดของชั้นภูมิ

PROC SURVEYREG DATA=Farms

TOTAL= StratumTotals;

STRATA State Region;

CLASS State Region;

MODEL CornYield = FarmArea State*Region

/SOLUTION CLPARM;

WEIGHT Weight;

ESTIMATE 'CornYield '

INTERCEPT 235 FarmArea 21950

State*Region 100 50 15 30 40;

RUN;

ภาพที่ 6 คำสั่ง SAS ที่ใช้ในการวิเคราะห์การถดถอยสำหรับ แผนแบบการชักตัวอย่างสุ่มแบบมีชั้นภูมิ

จากภาพที่ 6 มีชุดข้อมูลที่ชื่อ StratumTotals เป็นข้อมูลที่เก็บขนาดของชั้นภูมิ (ต่างจากการชักตัวอย่างสุ่มแบบง่ายที่บอกขนาดรวมของประชากรทั้งหมด) ส่วนคำสั่ง STRATA ใช้ระบุตัวแปรที่ใช้แบ่งชั้นภูมิซึ่งได้แก่รัฐและเขต ส่วนคำสั่ง CLASS เป็นการกำหนดชื่อตัวแปรแบ่งประเภทซึ่งในที่นี้คือตัวแปรที่ใช้ในการแบ่งชั้นภูมิ ให้สังเกตเทอม State*Region ในคำสั่ง MODEL ซึ่งเกิดจาก

การแบ่งชั้นภูมิ 5 ชั้น โดยโปรแกรมจะทำการสร้างตัวแปรคัมมี 4 ตัวดังตัวแบบ

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 D_{3i} + \beta_5 D_{4i} + \varepsilon_i$$

โดยที่ Y แทนตัวแปร CornYield และ X แทนตัวแปร FarmArea และ

$D_1 = 1$ ถ้าอยู่ในรัฐ Iowa เขต 1, และ $D_1 = 0$ กรณีอื่น

$D_2 = 1$ ถ้าอยู่ในรัฐ Iowa เขต 2, และ $D_2 = 0$ กรณีอื่น

$D_3 = 1$ ถ้าอยู่ในรัฐ Iowa เขต 3, และ $D_3 = 0$ กรณีอื่น

$D_4 = 1$ ถ้าอยู่ในรัฐ Nebraska เขต 1, และ $D_4 = 0$ กรณีอื่น

ถ้าใช้คำสั่ง CLASS จะต้องระบุคำสั่งย่อย SOLUTION เสมอ เพื่อให้พิมพ์ค่าประมาณของสัมประสิทธิ์การถดถอย ส่วนคำสั่ง WEIGHT ใช้ระบุตัวแปรชื่อ Weight เป็นตัวแปรที่ใช้เก็บค่าน้ำหนักของการชักตัวอย่าง สำหรับเทอม State*Region ในคำสั่ง ESTIMATE ใช้ระบุจำนวนประชากรไร่ข้าวโพดในแต่ละชั้นภูมิ และในการทำงานเดียวกันกับตัวอย่างที่ผ่านมาจะใช้วิธีตัวประมาณแบบถดถอยในการพยากรณ์ ผลลัพธ์ที่ได้จากคำสั่งข้างต้นบางส่วนแสดงในภาพที่ 7 และ 8

ANOVA for Dependent Variable CornYield					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	119862.2	23972.44	11.72	0.0002
Error	13	26582.0	2044.77		
Corrected Total	18	146444.2			
Tests of Model Effects					
Effect	Num DF	F Value	Pr > F		
Model	5	63.49	<.0001		
Intercept	1	2.78	0.1176		
FarmArea	1	96.45	<.0001		
State*Region	4	10.32	0.0004		

ภาพที่ 7 ตารางวิเคราะห์ความแปรปรวนและการทดสอบอิทธิพล ของเทอมต่างๆ ในตัวแบบ (จากคำสั่งในภาพที่ 6)

จากภาพที่ 7 ให้สังเกตที่ DF ของ Model ซึ่งมีค่าเท่ากับ 5 ซึ่งหมายถึง มีตัวแปรอิสระในตัวแบบทั้งหมด 5 ตัว (ซึ่งต่างจากการชักตัวอย่างสุ่มแบบง่ายที่มีตัวแปรอิสระเพียงตัวเดียว) และขนาดของพื้นที่เพาะปลูกมีผลต่อผลผลิตข้าวโพดจริง (p-value เท่ากับ 0.0002) นอกจากนี้ตัวแปร FarmArea และ State*Region ต่างก็มีความสำคัญ

Estimated Regression Coefficients						
	Standard	95% Confidence				
Parameter	Estimate	Error	t Value	Pr > t	Interval	
Intercept (b_0)	-13.276	5.340	-2.49	0.0262	-24.729	-1.823
FarmArea (b_1)	0.356	0.036	9.82	<.0001	0.278	0.434
State*Region Iowa 1						
(b_2)	18.269	9.220	1.98	0.0675	-1.506	38.044
State*Region Iowa 2						
(b_3)	33.474	8.308	4.03	0.0012	15.655	51.292
State*Region Iowa 3						
(b_4)	13.843	5.963	2.32	0.0359	1.054	26.631
State*Region Nebraska 1						
(b_5)	-31.208	11.140	-2.80	0.0141	-55.101	-7.316
State*Region Nebraska 2						
	0.000	0.000	.	.	0.000	0.000
Analysis of Estimable Functions						
	Standard	95% Confidence Parameter				
Estimate	Error	t Value	Pr > t	Interval		
CornYield	7463.523	926.842	8.05	<.0001	5475.646	9451.401

ภาพที่ 8 สัมประสิทธิ์การถดถอยและค่าพยากรณ์ผลผลิตข้าวโพด (จากคำสั่งในภาพที่ 6)

ผลลัพธ์ในภาพที่ 8 ค่าประมาณสัมประสิทธิ์การถดถอยของตัวแปร FarmArea เท่ากับ 0.356 เป็นค่าความชันที่เท่ากันทุกชั้นภูมิ โดยมีค่าตลาดเคลื่อนมาตรฐานของการประมาณค่าคือ 0.036 และในการทดสอบค่าสัมประสิทธิ์ β_2 พบว่าค่า p-value เท่ากับ 0.0675 ดังนั้นจึงสรุปว่าผลผลิตข้าวโพดของรัฐ Iowa เขตที่ 1 กับรัฐ Nebraska เขตที่ 2 ไม่แตกต่างกันอย่างมีนัยสำคัญ ส่วนค่าสัมประสิทธิ์อื่นๆ สามารถบอกความแตกต่างระหว่าง

ผลผลิตข้าวโพดในแต่ละเขตได้ เช่น รัฐ Iowa เขตที่ 2 มีผลผลิตข้าวโพดสูงกว่ารัฐ Iowa เขตที่ 1 โดยเฉลี่ยเท่ากับ $33.474 - 18.269 = 15.205$ ตัน และรัฐ Nebraska เขตที่ 1 มีผลผลิตข้าวโพดต่ำกว่ารัฐ Nebraska เขตที่ 2 โดยเฉลี่ยเท่ากับ 31.208 ตัน

นอกจากนี้ค่าประมาณผลผลิตข้าวโพด (โดยวิธีตัวประมาณแบบถดถอย) คือ 7,463.52 ตัน เมื่อทราบว่าเป็นพื้นที่ในการเพาะปลูกทั้งหมดเป็น 21,950 ไร่ และทราบจำนวนไร่ข้าวโพดในแต่ละชั้นภูมิเป็น 100, 50, 15, 30, และ 40 ไร่ ตามลำดับ หรืออยู่ในช่วง 5,475.65 ตัน ถึง 9,451.40 ตัน ด้วยความเชื่อมั่นร้อยละ 95 ค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่านี้คือ 926.84 ตัน

การใช้ PROC GLM และโปรแกรม SPSS

ในกรณีที่สมมติว่า ตัวอย่างเป็นตัวอย่างสุ่มจากประชากรที่มีขนาดอนันต์ต่างๆ ที่จริงๆ แล้วตัวอย่างนั้นถูกเลือกมาจากการใช้แผนแบบการชักตัวอย่างชนิดอื่นซึ่งเป็นการสำรวจจากประชากรที่มีขนาดจำกัด และใช้ PROC GLM หรือทำการวิเคราะห์ด้วยโปรแกรมสำเร็จรูป SPSS ผลลัพธ์ที่ได้จะแตกต่างกับการใช้ PROC SURVEYREG อย่างไรให้พิจารณาจากผลลัพธ์ที่ได้จากคำสั่งในภาพที่ 9

```
PROC GLM DATA = Farms;
  MODEL CornYield = FarmArea / CLPARM;
  ESTIMATE 'CornYield'
  INTERCEPT 235 FarmArea 21950;
RUN;
```

ภาพที่ 9 คำสั่งในการวิเคราะห์การถดถอยสำหรับตัวอย่างสุ่มจากประชากรที่มีขนาดอนันต์ด้วย

PROC GLM

ภาพที่ 10 แสดงผลลัพธ์จากการใช้ PROC GLM เพื่อวิเคราะห์การถดถอยสำหรับตัวอย่างสุ่มจากประชากรที่

มีขนาดอนันต์ และถ้าทำการวิเคราะห์ด้วยโปรแกรม SPSS โดยใช้เมนู Regression จะได้ผลลัพธ์ดังภาพที่ 11

The GLM Procedure					
Dependent Variable: CornYield					
Sum of					
Source	DF	Squares	Mean Square	F Value	Pr > F
Model	1	6388.903	6388.903	10.93	0.0042
Error	17	9939.097	584.653		
Corrected Total	18	16328.000			
		Standard		95%	
Parameter Estimate	Error	t Value	Pr > t	Confidence Limits	
Intercept	12.644	8.983	1.41	0.177	-6.308 31.596
FarmArea	0.184	0.056	3.31	0.0042	0.066 0.301
		Standard		95%	
Parameter Estimate	Error	t Value	Pr > t	Confidence Limits	
CornYield	7003.040	1376.079	5.09	<.0001	4099.768 9906.312

ภาพที่ 10 สัมประสิทธิ์การถดถอยและค่าพยากรณ์ผลผลิตข้าวโพดกรณีตัวอย่างสุ่มจากประชากรขนาดอนันต์

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6388.903	1	6388.903	10.928	.004 ^a
	Residual	9939.097	17	584.653		
	Total	16328.00	18			

a. Predictors: (Constant), FARMAREA

b. Dependent Variable: CORNYIELD

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error			
1	(Constant)	12.644	8.983		1.408	.177
	FARMAREA	.184	.056	.626	3.306	.004

a. Dependent Variable: CORNYIELD

ภาพที่ 11 การวิเคราะห์การถดถอยด้วยโปรแกรม SPSS

ตารางที่ 2 เปรียบเทียบค่าสัมประสิทธิ์การตัดสินใจ (R^2) ค่าประมาณความแปรปรวน (s) ค่าประมาณสัมประสิทธิ์การถดถอย (b_1) และ ค่าพยากรณ์ผลผลิตข้าวโพดและช่วงความเชื่อมั่นร้อยละ 95 กรณีต่างๆ

PROC (โปรแกรมSAS) หรือโปรแกรม SPSS	R^2	s	b_1 (SE)	ช่วงเชื่อมั่นของ b_1	ค่าพยากรณ์ (SE)	ช่วงเชื่อมั่นของ ค่าพยากรณ์
SURVEYREG						
- สุ่มแบบง่าย	39.1	24.1	0.184 (0.069)	(0.039,0.328)	7003.04 (1055.62)	(4785.26,9220.82)
- มีชั้นภูมิ	81.9	8	0.356 (0.036)	(0.278,0.434)	7463.52 (926.84)	(5475.65,9451.40)
		12.8				
		6				
GLM	39.1	24.1	0.184 (0.056)	(0.066,0.301)	7003.04 (1376.08)	(4099.77,9906.31)
		8				
โปรแกรม SPSS	39.1	24.1	0.184 (0.056)	(0.066,0.301)	7003.04 (1376.08)	(4099.77,9906.31)
		8				

3. บทสรุป

ตารางที่ 2 แสดงการเปรียบเทียบค่าสัมประสิทธิ์การตัดสินใจ (R^2), ค่าประมาณความแปรปรวน (s) และค่าประมาณสัมประสิทธิ์การถดถอยรวมทั้งค่าพยากรณ์ของผลผลิตข้าวโพดโดยวิธีตัวประมาณแบบถดถอยเมื่อทราบค่าพื้นที่ในการเพาะปลูกทั้งหมดเป็น 21,950 ไร่ ที่ได้จากคำสั่งต่างๆ ของโปรแกรม SAS และโปรแกรม SPSS ซึ่งจะเห็นว่าการใช้ PROC SURVEYREG กรณีตัวอย่างสุ่มแบบง่ายให้ค่า R^2 และ s ไม่แตกต่างจากการใช้ PROC GLM และการใช้เมนูคำสั่งในโปรแกรม SPSS ส่วนค่าประมาณสัมประสิทธิ์การถดถอยและค่าพยากรณ์ผลผลิตข้าวโพดนั้น คำสั่ง PROC SURVEYREG กรณีตัวอย่างสุ่มแบบง่าย ให้ค่าสัมประสิทธิ์และค่าพยากรณ์ไม่ต่างจาก PROC GLM และ โปรแกรม SPSS แตต่างกันที่ค่าคลาดเคลื่อนมาตรฐาน แต่ถ้าผู้วิจัยใช้แผนแบบการชักตัวอย่างแบบมีชั้นภูมิแล้วใช้ PROC GLM หรือโปรแกรม SPSS ค่าประมาณที่ได้จะไม่ถูกต้องและให้ค่าที่ต่ำกว่าค่าที่ควรจะเป็น และมีค่าคลาดเคลื่อนมาตรฐานของการ

ประมาณสูงกว่าส่งผลให้ช่วงเชื่อมั่นร้อยละ 95 ของการประมาณพารามิเตอร์ที่สนใจมีความกว้างมากกว่าช่วงเชื่อมั่นที่มีการปรับให้สอดคล้องกับแผนแบบการชักตัวอย่าง

การเลือกใช้คำสั่งในโปรแกรมสำเร็จรูปที่ถูกต้องนั้น นอกจากจะต้องให้สอดคล้องตามข้อตกลงเบื้องต้นของการใช้สถิติอื่นๆ แล้วยังจะต้องให้สอดคล้องกับแผนแบบการชักตัวอย่างที่เลือกใช้อีกด้วย โดยเฉพาะในทางปฏิบัติซึ่งเป็นการเลือกตัวอย่างจากประชากรที่มีขนาดจำกัด ซึ่งถ้าหากผู้ใช้ไม่ระวังในเรื่องนี้ ผลสรุปที่ได้จะมีความน่าเชื่อถือน้อยลงซึ่งอาจนำไปสู่การตัดสินใจที่ผิดพลาด โดยเฉพาะอย่างยิ่งหากการตัดสินใจในเรื่องนั้นๆ นำไปสู่นโยบายสำคัญ ที่มีผลต่อสังคมและเศรษฐกิจของชาติ

4. เอกสารอ้างอิง

- [1] What's New in Data Analysis : Sample Survey Design and Analysis,
<http://support.sas.com/rnd/app/da/new/dasurvey.html>, 2005. (January, 10, 2005)

- [2] Berglund, P. A., Analysis of Complex Sample Survey Data Using the SURVEYMEANS and SURVEYREG Procedures and Macro Coding. <http://www2.sas.com/proceedings/sugi27/p263-27.pdf> (January, 10, 2005)
- [3] The SURVEYLOGISTIC Procedure. http://support.sas.com/onlinedoc/913/getDoc/th/statug.hlp/surveylogistic_sect1.htm, 2003. (January, 10, 2005)
- [4] The SURVEYREG Procedure. http://support.sas.com/onlinedoc/913/getDoc/th/statug.hlp/surveyreg_index.htm, 2003 (January, 10, 2005)
- [5] An A., and Watts, D. New SAS Procedures for Analysis of Sample Survey Data., http://www.ats.ucla.edu/stat/sas/library/svy_survey.pdf (January, 10, 2005)