**เอกสารที่ 5**

# การจัดทำสถิติและการวิเคราะห์ข้อมูลภาคตัดขวาง
# (Cross section analysis)

**ภักดิ์ ทองส้ม**

**เสนอต่อสำนักงานส่งเสริมวิสาหกิจขนาดกลางและขนาดย่อม**

**30 กันยายน 2556**

# การจัดทำสถิติและการวิเคราะห์ข้อมูลภาคตัดขวาง(Cross section analysis)

ภักดิ์ ทองส้ม

30 กันยายน 2556

## 1. การจัดทำสถิติเพื่อใช้สำหรับงานวิจัย

### 1.1 ระเบียบวิธีวิจัย

การวิจัยทางสังคมศาสตร์หมายถึงกระบวนการการดำเนินงานเพื่อหาคำตอบของโจทย์ปัญหาทางสังคมศาสตร์(social science)ที่ต้องใช้ทั้งหลักวิชาการและวิธีการดำเนินงานทางสถิติที่ถูกต้องเพื่อให้สามารถสร้างความมั่นใจได้ว่าคำตอบที่ได้นั้นมีความเป็นจริง น่าเชื่อถือได้

1.1.1 ประเภทของงานวิจัย

ประเภทของงานวิจัยตามวิธีการทางสถิติมี 3 ประเภทคือ

1) การวิจัยเชิงประวัติศาสตร์ เช่นการค้นคว้าจากเอกสาร(desk research) ตัวอย่างในทางเศรษฐศาสตร์ คือการเก็บข้อมูล time series แล้วเอามาสร้างสมการหาความสัมพันธ์ของตัวแปรเป็นแบบจำลองทางเศรษฐศาสตร์เป็นต้น

2) การทดลอง ซึ่งในทางสังคมศาสตร์อาจทำได้ค่อนข้างจำกัดเพราะเป็นเรื่องของพฤติกรรม

3) การวิจัยเชิงพรรณนา(descriptive research) เป็นการวิจัยที่มุ่งหาคำตอบในปัจจุบันเป็นการเก็บข้อมูลแบบภาคตัดขวาง เพื่อนำมาวิเคราะห์หาคำตอบของเหตุและผลต่างๆ ซึ่งเป็นจุดมุ่งหมายของเอกสารฉบับนี้ที่ “ผู้รับจ้าง” ประสงค์ต้องการถ่ายทอดให้กับ ขว. ไว้ใช้ประโยชน์ในการทำงานวิจัยต่างๆ ต่อไป

1.1.2 ขั้นตอนหรือระเบียบวิธีของการดำเนินงานวิจัย ประกอบด้วย 7 ขั้นตอนคือ

1) การกำหนดปัญหา จะต้องเป็นโจทย์ปัญหาที่ใหญ่พอสมควร เช่น ในกรณีของการส่งเสริม SMEs ปัญหาดังกล่าวควรให้ผลกระทบในวงกว้างต่อการตัดสินใจกำหนดแนวทางการส่งเสริม เป็นต้น ต้องคุ้มค่ากับทรัพยากรที่ต้องใช้เพื่อดำเนินการค้นคว้าหาคำตอบโดยวิธีการวิจัยดังกล่าว

2) การกำหนดขอบเขตของการศึกษา เป็นขั้นตอนของการเอาประเด็นปัญหาและการศึกษามาเจอกัน เพราะการศึกษาจำเป็นต้องระบุขอบเขตว่าต้องการให้ครอบคลุมเฉพาะที่สนใจหรือสามารถดำเนินการได้

3) การกำหนดประชากรและกลุ่มตัวอย่าง เป็นการกำหนดหน่วยทางสถิติเพื่อทำการเก็บรวบรวมข้อมูล

4) กำหนดเครื่องมือที่ใช้เพื่อการรวบรวมข้อมูล เช่น การสังเกต การสัมภาษณ์ การส่งแบบบันทึก การเก็บข้อมูลทางไปรษณีย์หรือไปรษณีย์อีเล็กทรอนิกส์ เป็นต้น

5) การรวบรวมข้อมูล เป็นขั้นตอนที่มีรายละเอียดและข้อยุ่งยากพอสมควรโดยเฉพาะอย่างยิ่งการเก็บข้อมูลด้วยวิธีสัมภาษณ์ ขั้นตอนการรวบรวมข้อมูลคือการบริหารงานภาคสนามทั้งหมด

6) การประมวลผลและวิเคราะห์ข้อมูล

**7)** การจัดทำรายงานผลการศึกษา

**1.2 ประชากรและตัวอย่าง(population & sample)**

ประชากรหมายถึงจำนวนหน่วยของสถิติหรือข้อมูลที่มีอยู่ทั้งหมดตามที่ขอบเขตของการวิจัยที่กำหนดขึ้นมา ส่วนขอบเขตของประชากร(population frame) หมายถึงประชากรเฉพาะที่ต้องการศึกษาหรือสามารถอ้างอิงที่มาและทราบจำนวนได้อย่างชัดเจน ตัวอย่างเช่นจำนวนครัวเรือนทั้งหมดที่มีในพื้นที่จังหวัดกรุงเทพมหานครคือประชากร และจำนวนครัวเรือนที่มาขึ้นทะเบียนกับเขตต่างๆของกรุงเทพมหานครคือขอบเขตประชากรเป็นต้น ดังนั้นอาจมีบางครัวเรือนที่ตั้งอยู่ใน กทม. แต่ไม่ได้มาขึ้นทะเบียนจึงเป็นประชากรแต่ไม่ได้อยู่ในขอบเขต

ตัวอย่างหรือกลุ่มตัวอย่างหมายถึงจำนวนหน่วยของสถิติหรือข้อมูลในส่วนที่ผู้ทำการศึกษาวิจัยได้เลือกขึ้นมาเพื่อใช้เป็นตัวแทนของประชากรในขอบเขตที่กำหนด

**1.3 วิธีการสุ่มตัวอย่าง**

วิธีการสุ่มตัวอย่างหมายถึงการดำเนินการเพื่อให้ได้มาซึ่งตัวอย่างเพื่อทำการเก็บรวบรวมข้อมูลและใช้เป็นตัวแทนของประชากร

1.3.1 หลักเกณฑ์ของการเลือกตัวอย่าง ประกอบด้วย

1) สามารถใช้เป็นตัวแทนที่ดีของประชากรทั้งหมด ตัวอย่างเช่น ถ้าต้องการศึกษาเรื่องการไม่ได้รับการศึกษาของเด็กวัยเรียนใน กทม. ถ้าใช้ครัวเรือนจดทะเบียนในแต่ละเขตเป็นตัวอย่าง ก็อาจไม่สามารถเป็นตัวแทนที่ดีของประชากร

2) มีขนาดที่เหมาะสมเพื่อให้เกิดความคลาดเคลื่อน แต่ก็จำเป็นที่ต้องอยู่ภายใต้งบประมาณ กำลังคนและกรอบเวลาเวลาที่กำหนดในการศึกษา

3) ใช้วิธีการสุ่มตัวอย่างที่ถูกต้องตามหลักวิธีการทางสถิติ

1.3.2 แนวคิดของการกำหนดขนาดตัวอย่าง

หลักการของกำหนดขนาดของตัวอย่างควรคำนึงถึงการใช้จำนวนตามวิธีการของสถิติเป็นความสำคัญลำดับแรกเพื่อลดความคลาดเคลื่อนให้น้อยที่สุด ซึ่งจะทำให้ผลการศึกษาที่ได้ก็จะถูกต้องและใช้ประโยชน์ได้ แต่มิได้หมายความว่าถ้าใช้ตัวอย่างในจำนวนมากแล้วทำให้ไม่มีความคลาดเคลื่อน เพราะขึ้นอยู่กับวิธีการที่ให้ได้มาซึ่งตัวอย่าง จำนวนตัวอย่างถ้ามีเพียงพอ(ต่อขนาดของประชากร)แล้วนั้น ถ้าหากจะเพิ่ม(หรือลด) ตัวอย่างลงก็อาจจะไม่มีผลกระทบต่อ error มากนักเมื่อเทียบกับจำนวนตัวอย่างที่มีอยู่ในจำนวนน้อย หลังจากนั้นลำดับต่อไปจึงจะพิจารณาว่าสามารถมีงบประมาณดำเนินได้มากน้อยเพียงใด ที่พบบ่อยครั้งคืองบประมารไม่เพียงพอ ถ้ามีผลกระทบไม่มากนักต่อจะนวนตัวอย่างภายใต้ของเขตที่ศึกษาก็อาจดำเนินการได้ แต่ถ้ามีผลกระทบมากอาจจำเป็นต้องกำหนดของเขตของการศึกษาใหม่ให้เหมาะสม หรือมุ่งเน้นเฉพาะตัวอย่างที่ต้องการศึกษาโดยแท้จริงเท่านั้น

1.3.3 วิธีการสุ่มตัวอย่าง

การสุ่มตัวอย่างมี 2 วิธีคือการสุ่มแบบใช้หลักของความน่าจะเป็น(probability) และการสุ่มแบบที่ไม่ได้ใช้หลักความน่าจะเป็น

การสุ่มแบบใช้หลักของความน่าจะเป็นแบ่งออกไดเป็น 4 ประเภท

1) Simple random sampling หมายถึงวิธีการสุ่มที่ตัวอย่างทั้งหมดมีโอกาสถูกเลือกมาที่เท่าๆกัน เช่น การจับสลาก การสุ่มวิธีนี้มีข้อดีคือง่ายจึงเป็นที่นิยม วิธีการคือเพียงแค่เอารายชื่อประชากรทั้งหมดมาใช้และให้หมายเลขเท่านั้น แต่วิธีนี้ควรใช้กับประขากรที่มีจำนวนไม่มากนัก เครื่องมือที่นิยมใช้กับวิธีนี้คือตาราง random number

2) Systematic random sampling เป็นการสุ่มเหมือน simple random sampling แต่มีการคำนวณหาระยะห่างระหว่างตัวอย่าง โดยเอาจำนวนประชากรหารด้วยตัวอย่าง หลังจากนั้นสุ่มตัวอย่างแรกส่วนตัวอย่างต่อไปก็นับตามระห่างดังกล่าว

3) Stratified random sampling เป็นวิธีที่ใช้กับประชากรขนาดใหญ่ การสำรวจของสำนักงานสถิติแห่งชาตอที่ครอบคลุมตัวอย่างทั่วประเทศจะใช้วิธีการสุ่มวิธีนี้ โดยสามารถจัดชั้นตัวอย่างได้หลายระดับ(stage)หรือชั้นต่างๆ เช่น 2-stage, 3-stage เป็นต้น หลักการคือต้องให้ตัวอย่างในแต่ละชั้นมีลักษณะเหมือนกันมากที่สุด

4) Cluster random sampling เป็นการสุ่มแบบแบ่งกลุ่มย่อยเหมือนกับ stratified random sampling แต่ที่ต่างกันที่สำคัญคือตัวอย่างในแต่ละกลุ่มต้องมีความแตกต่างกันให้น้อยที่สุด หรือผสมผสานกันมากที่สุด

ส่วนวิธีการสุ่มที่ไม่ได้ใช้หลักความน่าจะเป็นนั้น โดยหลักการแล้วควรหลีกเลี่ยงที่จะนำมาใช้ ถ้าจำเป็นควรใช้ด้วยความระมัดระวัง รวมทั้งการตีความของค่าที่ได้จากตัวอย่างด้วยเช่นกัน การสุ่มแบบที่ไม่ได้ใช้หลักความน่าจะเป็นมีย่อย ๆ หลายวิธี แต่วิธีที่นิยมกันคือ purposive sampling หรือการสุ่มแบบเจาะจง

1.3.4 วิธีการกำหนดขนาดตัวอย่างมี 2 วิธี

1.3.4.1 ใช้วิธีการคำนวณเพื่อกำหนดจำนวนตัวอย่าง

ในกรณีทราบจำนวนประชากร จะใช้สูตรการคำนวณดังนี้

$$n = N/(1+N(e)^2)$$

เมื่อ n คือจำนวนตัวอย่างที่ต้องใช้

N คือจำนวนประชากรที่ทราบจำนวน

e คือ % ความคลาดเคลื่อนที่ยอมรับได้ เช่น 5% ดังนั้น e = .05

ในกรณีไม่ทราบจำนวนประชาการ ใช้สูตรการคำนวณดังนี้

$$n = (P(1-P)(Z)^2)/ (e)^2$$

เมื่อ n คือจำนวนตัวอย่างที่ต้องใช้

P คือ % ที่ต้องการจะสุ่มจากประชากรทั้งหมด

Z คือค่ามาตรฐานหรือ Z score ซึ่งควรใช้ที่ 95%(หรือมากกว่าขึ้นไป) ดังนั้น z= 1.96

e คือ % ความคลาดเคลื่อนที่ยอมรับได้ เช่น 5% ดังนั้น e = .05

1.3.4.2 ใช้ค่าจากตารางสำเร็จรูปที่นักสถิติได้มีการคำนวณไว้ ที่มีการใช้กันมากคือค่าของ Taro Yamane

นอกจาก 2วิธีดังกล่าวแล้ว ยังสามารถใช้โปรแกรมคอมพิวเตอร์สำเร็จรูปทางสถิติคำนวณหาจำนวนตัวอย่างได้ด้วยเช่นกัน

**1.4 แบบสอบถาม**

การได้มาซึ่งข้อมูลที่ผู้วิจัยต้องการสามารถใช้เครื่องมือต่าง ๆได้หลายรูปแบบ เช่นการทดสอบ การสัมภาษณ์หรือการสังเกตซึ่งไม่ว่าวิธีใดต้องมีแบบฟอร์มสำหรับบันทึกข้อมูลตามที่ต้องการด้วย เช่น การทดสอบโดยใช้แบบทดสอบ การสัมภาษณ์โดยใช้แบบฟอร์มสัมภาษณ์(เช่นการสัมภาษณ์เพื่อคัดเลือกพนักงานใหม่) หรือการสังเกตซึ่งต้องมีแบบฟอร์มสำหรับกรอกข้อมูลโดยผู้ที่ทำการสังเกตด้วย หรือเก็บ

ข้อมูลจากการบันทึกประจำวัน เช่นการให้แบบฟอร์มให้แม่บ้านบันทึกรายการอาหารที่ครัวเรือนบริโภคประจำวันในแต่ละวันรอบสัปดาห์ หรือการใช้วิธีสนทนากลุ่มเป็นต้น รวมทั้งการสอบถามโดยใช้แบบสอบถาม

แบบสอบถามคือเครื่องมือวิจัยประเภทหนึ่งเพื่อรวบรวมข้อมูลที่เป็น “ความจริง” จากผู้ถูกสอบถามหรือตัวอย่างหรือหน่วยสถิติ ลักษณะคำถามในแบบสอบถามจำแนกได้ 2 ประเภทคือคำถามแบบเปิดและคำถามแบบปิด คำถามแบบเปิดมีข้อดีคือให้อิสระเต็มที่กับผู้ตอบเต็มที่ จึงควรใช้กับการถามประเภทความเห็น แต่ในการประมวลผลผู้วิจัยต้องรวบรวมและจัดเป็นกลุ่มๆ ของคำตอบที่อยู่ในประเภทเดียวกันเข้าด้วยกัน(เคยปรากฏมีที่ปรึกษาที่ประมวลคำตอบแบบปลายเปิดโดยการรวบรวมคำตอบทั้งหมดทุกๆแบบส่งมาให้) ส่วนคำถามแบบปิดเป็นวิธีให้ผู้ตอบเลือกได้เฉพาะที่กำหนดมาให้เท่านั้น คำถามประเภทนี้สามารถแบ่งย่อยออกได้เป็น 3 ประเภทคือ 1) check-list, 2) ranking, 3) rating

คุณลักษณะที่สำคัญของแบบสอบถามคือต้องสั้น กระชับ ใช้ถ้อยคำที่เข้าใจง่าย และควรเรียงลำดับจากคำตอบง่ายๆ ไปหายาก โครงสร้างของแบบสอบถาม โดยทั่วไปจะประกอบด้วย 3 ส่วนคือ

1) ข้อมูลทั่วไปของตัวอย่างหรือผู้ตอบแบบสอบถาม
2) ข้อถามความจริงหรือข้อมูลหรือตัวเลขต่างๆที่ต้องการรวบรวมข้อมูลสำหรับการศึกษาวิเคราะห์
3) ความเห็นหรือทัศนคติที่มีต่อปัญหาที่นักวิจัยต้องการศึกษา

ทั้งนี้ข้อถามส่วนแรกและส่วนที่สองควรสัมพันธ์กัน ข้อมูลส่วนแรกควรเป็นปัจจัยหรือเหตุผลที่สามารถใช้อธิบาย facts ที่เกดขึ้นในส่วนที่สองได้

หลักการของวิธีการออกแบบโครงสร้างของแบบสอบถามมี 7 ขั้นตอน

1) ศึกษา รวบรวมข้อมูลเบื้องต้นทั่วไปเกี่ยวกับปัญหาที่ต้องการวิจัย เพื่อมากำหนดโครงสร้างคำถาม หรืออาจใช้ประสบการณ์ องค์ความรู้ในเรื่องนั้นๆ หรือดูจากงานศึกษาเก่าๆที่คล้ายคลึงกัน
2) กำหนดโครงสร้างของแบบสอบถาม กำหนดคำที่สำคัญๆ นิยามและสำนวนภาษา
3) ยกร่างข้อคำถาม กำหนดว่าคำถามแต่ละข้อควรเป็นแบบใด(เปิด, ปิด, check-list, ranking หรือ rating เป็นขั้นตอนที่ได้แบบสอบถามในเบื้องต้น
4) การขอความเห็นจากผู้รู้หรือผู้เชี่ยวชาญ
5) ทดสอบแบบสอบถามกับตัวอย่างจริง ค้นหาว่าผู้ที่จะตอบแบบสอบถามมีความเข้าใจตรงกับที่นักวิจัยต้องการถามหรือไม่ ตรวจสอบว่าถ้าไม่มีการอธิบายใดๆเพิ่มเติมผู้ตอบจะเข้าใจคำถามหรือไม่ โดยหลักการแล้วการทดสอบแบบสอบถามมิใช่เป็นการเก็บข้อมูลเพื่อนำมาใช้วิจัย แต่เป็นการค้นหาความถูกต้องเหมาะสมของแบบสอบถามกับหน่วยสถิติหรือผู้ตอบที่อยู่ในกลุ่มประชากร ถ้านักวิจัยเห็นว่าแต่ละกลุ่มมีความแตกต่างกันควรทดสอบแยกแต่ละกลุ่มของ

ตัวอย่าง จำนวนของการทดสอบอาจไม่จำเป็นที่ต้องมีจำนวนมาก ขึ้นอยู่กับข้อความจริงที่ได้ (พบว่ามีงานวิจัยที่ที่ปรึกษาดำเนินการให้กับ สสว. ได้เอาคำตอบที่ได้จากขั้นตอนทดสอบแบบสอบถามมาใช้จริงซึ่งต้องใช้อย่างระมัดระวัง)

6) ปรับปรุงแบบสอบถามให้เหมาะสม(ในขั้นตอนนี้มีนักสถิติบางท่านเสนอว่าควรมีการทดสอบแบบสอบถามซ้ำอีกครั้ง แต่อาจไม่จำเป็นมากนัก)
7) นำแบบสอบถามไปใช้งานจริง

## 1.5 วิธีการเก็บข้อมูล

วิธีการเก็บข้อมูลมีหลายวิธี ที่สำคัญได้แก่

1) การส่งทางไปรษณีย์(จากประสบการณ์พบว่าวิธีนี้ได้ผลค่อนข้างน้อย ตัวอย่างเช่นในกรณีของแบบสอบถามที่ง่ายที่สุดโดยถามความเห็นผู้ประกอบการธุรกิจและคำตอบเป็นลักษณะ check list ความยาวประมาณ 1 หน้า เป็นไปรษณีย์ตอบกลับโดยไม่ต้องปิดแสตมป์ พบว่าอัตราได้รับแบบคืนต่ำประมาณ 5 % แต่ถ้าผู้ตอบเป็นครัวเรือนคาดว่าน่าที่จะสูงกว่านี้)
2) การสอบถามทางโทรศัพท์
3) การสัมภาษณ์โดยตรง(face to face) ซึ่งมีทั้งโดยวิธีทอดแบบ(ส่งแบบสอบถามให้ศึกษาล่วงหน้าแล้วนัดวันมาเก็บแบบหรือสัมภาษณ์) และไม่มีการทอดแบบ วิธีนี้เป็นวิธีที่คาดหวังผลได้ดีที่สุด
4) การสอบถามทาง internet

สิ่งที่จะต้องส่งให้กับผู้ตอบแบบร่วมไปกับแบบสอบถาม ขึ้นอยู่กับวิธีการเก็บข้อมูล เช่นวิธีสัมภาษณ์ หรือการส่งทางไปรษณีย์ โดยทั่วไปแล้วควรประกอบด้วย 1)หนังสือนำ 2) คำอธิบายข้อคำถาม และ 3)แบบสอบถาม โดยจะต้องมีคำระบุการเก็บข้อมูลต่างๆ เป็นความลับ ไม่มีการเผยแพร่เป็นรายบุคคล ไว้ด้วย

## 1.6 การบริหารจัดการงานภาคสนาม

โดยแท้จริงแล้วงานบริหารงานภาคสนามมีขอบเขตงานที่ครอบคลุมขั้นตอนของการดำเนินงานภาคสนามทั้งหมด เป็นขั้นตอนที่ยุ่งยาก แต่มีความสำคัญค่อนข้างมากต่อการรวบรวมข้อมูลโดยวิธีการสัมภาษณ์ผู้ตอบแบบสอบถามโดยตรง เรื่องที่สำคัญๆ ที่ต้องพิจาณาในขั้นตอนของภาคสนามที่สำคัญคือ

1) การจัดเตรียมบุคลากรทั้งพนักงานสัมภาษณ์(enumerators) พนักงานพี่เลี้ยง(supervisors) ต้องทำการสรรหา การอบรมให้ความรู้ กำหนดหน้าที่ให้ชัดเจน
2) การประสานงาน ต้องดำเนินการล่วงหน้าและควรใช้บุคลากรในพื้นที่ดำเนินการ ในกรณีการสอบถามผู้ประกอบการควรประสานงานกับ อุตสาหกรรมจังหวัด พาณิชย์จังหวัด เป็นต้น ใน

บางกรณีอาจจำเป็นต้องให้หน่วยงานดังกล่าวออกหนังสือแนะนำตัวไปด้วย หรือมีเจ้าหน้าที่ของหน่วยงานดังกล่าวร่วมไปสัมภาษณ์ด้วย

3) การสัมภาษณ์ ควรนัดล่วงหน้ากับตัวอย่าง พนักงานสัมภาษณ์ต้องแต่งตัวให้เรียบร้อย มีบัตรประจำตัวแสดงให้เห็นชัดเจน มีการแนะนำตัว การใช้วาจาสุภาพ การสัมภาษณ์ควรเป็นการซักถามตัวต่อตัว บุคคลที่ไม่เกี่ยวข้องไม่ควรอยู่ในขณะสอบถาม อธิบายวัตถุประสงค์ของงาน ประโยชน์ที่จะเกิดขึ้น การถามไม่จำเป็นต้องเรียงคำถามตามแบบสอบถาม ควรเรียงจากง่ายไปหายาก สอบถามเสมือนเป็นการพูดคุยปรกติ(ดังนั้นพนักงานภาคสนามต้องจดจำคำถามในแบบให้ได้แม่ยำ)หลังจากนั้นจึงดูว่ายังมีส่วนใดที่ยังไม่มีคำตอบ ก็ถามคำถามเฉพาะส่วนนั้นๆ ในบางช่วงควรให้ผู้ตอบได้แสดงความคิดเห็นบ้างเพื่อเป็นการผ่อนคลาย การถามควรถามตั้งแต่ต้นจนจบ ไม่ควรมี break และไม่ควรใช้เวลายาวนานมากเกินไป
4) พนักงานสนามต้องทำหน้าที่ตรวจสอบคำตอบต่างๆ (edit) ในแบบสอบถามให้ครบถ้วน ถูกต้องในขั้นตอนของภาคสนาม ก่อนที่จะส่งให้ผู้รวบรวมจึงติดต่อกันไป
5) เจ้าหน้าที่ Supervisor จะต้องทำหน้าที่ทั้งการให้คำปรึกษาแก่ enumerator และแก้ปัญหาต่างๆ ในระดับพื้นที่ทั้งหมด
6) ระยะเวลาการเก็บข้อมูลควรเป็นเฉพาะเวลากลางวัน(พระอาทิตย์ขึ้นถึงพระอาทิตย์ตกซึ่งเป็นเวลาที่กฎหมายสถิติของประเทศไทยกำหนดไว้)
7) ในกรณีจำเป็นที่ต้องมีการซ่อมแบบ(แบบที่เก็บข้อมูลมาได้คำตอบที่ไม่สมบูรณ์ ต้องเก็บเพิ่มเติมส่วนที่ขาดหายไป) ควรดำเนินการทันที

**1.7 การกำหนดระยะเวลาดำเนินงาน**

การกำหนดระยะเวลาในขั้นตอนของการเก็บข้อมูลควรเป็นระยะเวลาที่สั้น เพื่อป้องกันมิให้ปัจจัยที่อาจเกิดผลกระทบต่อตัวอย่างแต่ละรายแตกต่างกัน

ถ้าจำนวนพื้นที่ที่ต้องดำเนินการเก็บข้อมูลมีมากหลายพื้นที่ เช่นครอบคลุมทั่วประเทศ ในกรณีนี้ต้องดำเนินการในทุกๆพื้นที่ไปพร้อมๆ กัน

ในบางกรณีอาจเกิดเหตุการณ์สำคัญ เช่นเกิดภัยพิบัติน้ำท่วม ไม่สามารถเข้าพื้นที่ที่ได้รับผลกระทบนั้นได้ ในกรณีนี้อาจจำเป็นต้องตัดพื้นที่ดังกล่าวออกไป หรือถ้าจะดำเนินการต่อเมื่อสามารถเข้าพื้นที่ได้แล้วคำถามและคำตอบถามจะต้องเป็นการย้อนหลังให้ตกอยู่ในช่วงเวลาที่สำรวจภาคสนามปรกติ(ในทางปฏิบัติจริงกรณีเกิดมหาอุทกภัยปี 2554 ที่ผ่านมาพบว่าผู้ตอบแบบสอบถามก็ยังคงตอบจากความรู้สึกบนผลกระทบของอุทกภัยแม้ว่าจะขอให้ตอบในช่วงระยะเวลาที่ปรกติก็ตาม)

**1.8 การประมวลข้อมูล**

การประมวลผลคือการรวบรวมข้อมูลจากหน่วยสถิติแต่ละรายให้เป็นภาพรวมของตัวอย่างและประชากรทั้งหมด ส่วนการวิเคราะห์เป็นการหาค่าทางสถิติต่างๆ แบ่งออกเป็นการวิเคราะห์เชิงพรรณนา

(descriptive) ได้แก่การหาค่าทางสถิติเช่น ค่ากลาง ค่าการกระจาย การเบี่ยงเบน ค่าความสัมพันธ์หรือสหสัมพันธ์ และการวิเคราะห์เชิงอนุมานเป็นการตีความจากตัวอย่างไปสู่ประชากรซึ่งเกี่ยวข้องกับทฤษฎีความน่าจะเป็นและการทดสอบสมมุติฐานทั้งหมดๆ

ขั้นตอนของการประมวลข้อมูล ประกอบด้วย

1) การตรวจสอบข้อมูล
2) การลงรหัส
3) การ key ข้อมูลลงโปรแกรมประมวลผลหรือโปรแกรมฐานข้อมูล
4) การประมวลข้อมูล
5) การตรวจสอบผลที่ประมวลได้

**1.9 การวิเคราะห์และรายงานผลการศึกษา**

เป็นขั้นตอนของการนำข้อมูลที่เก็บรวบรวมได้มาทำการวิเคราะห์ เป้าหมายของการวิเคราะห์คือการค้นหาคำตอบของโจทย์งานวิจัยที่ได้กำหนดไว้ ในกรณีงานวิจัยประเภทเชิงพรรณนาขั้นตอนนี้ก็คือการวิเคราะห์ข้อมูลภาคตัดขวาง(cross section analysis)

หลังจากนั้นเป็นการนำผลที่ได้จาการวิเคราะห์มาจัดทำรายงานเผยแพร่ การนำเสนอข้อมูลในรายงานสามารถแสดงเป็นได้ทั้งตาราง รูปกราฟ และเชิงพรรณนา ประกอบกันไป การเขียนรายงานควรเขียนให้ครอบคลุมข้อมูลที่ได้มาจากการวิเคราะห์ทั้งหมดให้ครบถ้วน ตรงตามประเด็นแต่ละประเด็น และสามารถอธิบายโจทย์ปัญหาได้ถูกต้อง ชัดเจน คลอบคลุมตามวัตถุประสงค์ที่กำหนดไว้ เนื้อหาของรายงานอย่างน้อยควรประกอบด้วย 5 ส่วนคือ

1) บทนำ ระบุปัญหาคืออะไร มีความสำคัญอย่างไร เหตุผลและความจำเป็นที่ต้องวิจัย วัตถุประสงค์ ขอบเขตงาน ตลอดจนนิยามศัพท์ ข้อกำหนด เงื่อนไขและข้อจำกัดต่างๆ
2) ทบทวนงานวิจัยที่เกี่ยวข้อง ผลการศึกษาที่เกิดขึ้นก่อนหน้า
3) ระเบียบวิธีวิจัย
4) ผลการศึกษาที่ได้(สามารถแยกประเด็นย่อยได้ ถ้ามีเนื้อหามาก)
5) สรุปและข้อเสนอแนะ

**2. การวิเคราะห์ข้อมูลภาคตัดขวาง(cross section analysis)[1]**

[1] เนื้อหาสาระของ cross section analysis ในส่วนนี้ทั้งหมดเหมือนกับบทสรุปในเรื่องสถิติที่นำเสนอในส่วนของหัวข้อเศรษฐมิติ แบบจำลองเศรษฐกิจมหภาค(macromodel) การนำมานำเสนอในส่วนนี้อีกครั้งเพื่อให้เอกสารของการจัดทำสถิติและการวิเคราะห์ข้อมูลภาคตัดขวางมความครบถ้วน

การวิเคราะห์ข้อมูลภาคตัดขวางมี 2 ระดับ ระดับแรกเป็นการวิเคราะห์เบื้องต้น บนข้อมูลแต่ละรายการที่เก็บรวบรวมมา หรืออาจเรียกได้ว่าเป็นการหาค่าสถิติต่างๆ จากข้อมูลก็ได้เช่นกัน ค่าที่หาเช่น ความถี่หรือการกระจาย ค่ากลาง ค่าเบี่ยงเบน เป็นต้น ส่วนการวิเคราะห์ระดับสองเป็นการวิเคราะห์ที่ลึกลงไป มีความยุ่งยากและรายละเอียดขึ้น เป็นการวิเคราะห์ถึงความสัมพันธ์ระหว่างตัวแปรหนึ่งกับอีกตัวแปรหนึ่งซึ่งโดยวิธีที่ง่ายที่สุดคือการดูค่าสหสัมพันธ์ แต่ค่าดังกล่าวจะบอกได้เพียงว่าตัวแปรทั้งสองมีความสัมพันธ์กันหรือไม่ และในทิศทางใด(ดูจากเครื่องหมาย) ไม่สามารถบอกได้ว่าตัวแปรอะไรเป็นเหตุอะไรเป็นผล ส่วนความสัมพันธ์จากการวิเคราะห์อีกวิธีหนึ่งคือการสร้างสมการถดถอยเชิงเส้นตรง และถ้าความสัมพันธ์ดังกล่าวมีทฤษฎีทางเศรษฐศาสตร์กำกับ สามารถระบุได้ว่าตัวแปรอะไรเป็นเหตุ อะไรเป็นผล ความสัมพันธ์ควรเป็นแบบทิศทางใด สมการที่ได้ก็จะเป็นแบบจำลองทางเศรษฐกิจ(economic model)

**2.1 Frequency distribution**

จากข้อมูลที่ได้มาในเบื้องต้นต้องพิจารณาว่าเป็นข้อมูลดิบแต่ละรายรวมเข้าด้วยกัน(ungrouped data) หรือเป็นข้อมูลที่ได้มีการจัดกลุ่มไว้แล้ว(grouped data) การจัดข้อมูลให้เป็น grouped data มีความสำคัญและจำเป็นในกรณีที่ถ้ามีข้อมูลเป็นจำนวนมาก สูตรการคำนวณต่างๆ ในที่นี้เป็นแบบ ungrouped data แต่ก็สามารถนำไปใช้กับ grouped data ได้เช่นกัน

ค่าสถิติพื้นฐานของการคำนวณหา frequency ที่สำคัญมีดังนี้

2.1.1 Absolute frequency

Absolute frequency = the number of observations in each class

2.1.2 Relative frequency

Relative frequency = dividing the number of observation in each class by the total number of observations in the data as a whole

2.1.3 Histogram

Histogram = a bar graph of a frequency distribution

2.1.4 Frequency polygon

Frequency polygon = a line graph of a frequency distribution

2.1.5 Cumulative frequency distribution

Cumulative frequency distribution = the total number 0f observations in all classes up and including the class

2.1.6 Ogive

Ogive = a distribution curve

**2.2 Measure of central tendency**

2.2.1 Mean(the arithmetic mean or average)[2]

Ungrouped data

$$\mu = (\sum X)/N$$

Grouped data

$\mu = (\sum fx)/N$ , เมื่อ $f$ = frequency of each class, x = the class midpoint

2.2.2 Median

Ungrouped data

Median = the ((N+1)/2)

Grouped data

Median = $L + ((n/2 - F)/f_m) * c$

L = lower limit of the median class

n = the number of observation in the data set

F = sum of the frequencies up to but not including the median class

$f_m$ = frequency of the median class

c = width of the class interval

2.2.3 Mode

Ungrouped data

Mode = the value that occurs most frequently in the data set

Grouped data

Mode = L + (d1/(d1 + d2)) * c

L = lower limit if the modal class(i.e., the class with the greatest frequency)

d1 = frequency of the modal class minus the frequency of the previous class

d2 = frequency of the modal class minus the frequency of the following class

c = width of the class interval

**2.3 Measures of dispersion**

2.3.1 Average deviation(AD)

---

[2] Mean หรือ arithmetic mean หรือ average คือค่า central tendency ที่นิยมใช้กันมากที่สุด ข้อจำกัดของ mean คือ จะถูกกระทบด้วยค่า extreme value ในข้อมูลชุดนั้นๆ การวัดค่า central tendency อื่นๆ อีก เช่น weighted mean, geometric mean harmonic mean เป็นต้น

$$AD = (\sum | x-\mu |)/N$$

2.3.2 Variance

$$\sigma^2 = (\sum (x-\mu)^2) /N$$

2.3.3 Standard deviation

$$\sigma = \sqrt{(\sum (x-\mu)^2)} /N$$

3 Coefficient of variance($V$)

$$V = \sigma / \mu$$

หมายเหตุ $\sigma$ for population, s = for sample

$\mu$ for population, x = for sample

N for population, n = for sample

**2.4 Shape of frequency distribution**

2.4.1 Skewness

Zero skewness = symmetrical about its mean

Positively skewness = right tail is longer, mean > median > mode

Negatively skewness = left tail is longer, mean , < median < mode

Skewness = the Pearson's coefficient of skewness, the third moment

2.4.2 Kurtosis

Leptokurtic = peaked curve

Platykurtic = flat curve

Kurtosis = the fourth moment

**2.5 การวิเคราะห์สถิติอนุมาน(inference statistics)**

สาระสำคัญของสถิติอนุมานคือการศึกษาเรื่องโอกาสความน่าจะเป็น(probability) ถ้าสมมุติว่าเหตุการณ์ A สามารถเกิดขึ้นได้ $n_A$ ครั้ง จากจำนวนเหตุการณ์ทั้งหมด $N$ ครั้งที่เป็นไปได้และทุกๆครั้งมีโอกาสที่เท่าเทียมกัน ดังนั้น ความน่าจะเป็นที่จะเกิด A, แสดงโดย $P$ (A) มีค่าเท่ากับ $n_A / N$(การแสดงค่า probability สามารถแสดงได้อีกทางหนึ่งโดยรูปภาพ เรียกว่า Venn diagram)

ค่า probability มี 2 ลักษณะคือ probability of single event และ probability of multiple event ซึ่งในกรณีหลังนี้ยังแบ่งออกเป็น mutually exclusive events และ not mutually exclusive events โดยมีกฎพื้นฐานเกี่ยวกับ multiple event probability ที่สำคัญ คือ

1) Rule of addition for mutually exclusive events

$P$ (A or B) = $P$ (A) + $P$ (B)

2) Rule of addition for not mutually exclusive events

$P$ (A or B) = $P$ (A) + $P$ (B) - $P$ (A and B)

3) Rule of multiplication for independent events

$P$ (A and B) = $P$ (A) * $P$ (B)

4) Rule of multiplication for dependent events

$P$ (A and B) = $P$ (A) * $P$ (B / A)

$P$ (B / A) = conditional probability of B, given that A has already occurred

2.5.1 Random variable

Random variable หมายถึงตัวแปรที่ค่าของตัวแปรนั้นมีความสัมพันธ์กับโอกาสที่จะเกิดขึ้น (A random variable is a variable whose value are associated with some probability of being observed)

ตัวแปร X เป็น random variable ถ้าหากทุกๆ ค่าจริง(real number) ซึ่งเท่ากับ a ใดๆแล้ว ตัวแปร X มีค่า probability ที่จะมีค่าเท่ากับหรือน้อยกว่า a, เขียนเป็นสัญลักษณ์ได้ว่า

$P (X \leq a)$

2.5.2 Probability distribution

หมายถึงรูปแบบ(formula) ของความน่าจะเป็น ในการให้ได้มาซึ่งค่าของ random variables ในค่าต่างๆ

random variable มี 2 รูปแบบ คือ discrete random variable และ continuous random variable, รูปแบบหนึ่งของ discrete random variable ที่สำคัญคือ binomial distribution, รวมทั้งการหาค่า mean, standard deviation ของ binomial distribution ดังกล่าว

2.5.3 Normal probability distribution

หมายถึง continuous random variable ที่มีรูปแบบการกระจายของ probability ที่จะเกิดค่า X ขึ้นในระหว่าง(interval) a และ b, มีรูปเป็นระฆังคว่ำ (bell-shaped distribution) การกระจายในลักษณะ normal probability distribution มีความสำคัญและใช้ค่อนข้างมากในทางสถิติ โดยทั่วไปจะใช้โดยระบุในทำนองว่า "เมื่อ X มีการกระจายเป็น normal distribution ซึ่งมี mean เท่ากับ $\mu$ และ variance เท่ากับ $\sigma^2$ จะเขียนแทนด้วยว่า

$$X \sim N(\mu, \sigma^2)$$

The central limit theorem = เมื่อขนาดของตัวอย่างมีจำนวนเพิ่มขึ้น($n \rightarrow \infty$) การกระจายตัวของตัวอย่างจะเป็น normal distribition โดยไม่คำนึงว่า parent population จะมีรูปร่างอย่างไร การใช้ theorem ดังกล่าวนี้ได้เมื่อ n > 30 และสามารถคำนวณหาค่า Z

$$Z = (X - \mu) / \sigma$$

2.5.4 $\chi^2$-distribution, *t*-distribution และ *F*-distribution

เป็นรูปแบบการกระจายอื่นๆ ที่นอกเหนือจาก normal distribution ที่มีการใช้กันค่อนข้างมากในทางสถิติ

$\chi^2$-distribution หมายถึงรูปแบบการกระจายของค่า sum of square of *n* independent standard normal variables, $Z \sim \chi^2$, เมื่อ $Z = \sum (x^2)$, โดยที่ $x_i \sim IN(0,1)$ ทั้งนี้ I = independent, N = normal

*t*-distribution หมายถึงการกระจายของ $Z = x / \sqrt{(y/n)}$ เมื่อ x และ y เป็นตัวแปรที่เป็นอิสระต่อกัน โดยที่ x มีการกระจายเป็น normal, mean = 0, variance = 1 และ y มีการกระจายเป็น $\chi^2$

*F*-distribution หมายถึงการกระจายของ $Z = (y_1/n_1) / (y_2/n_2)$ เมื่อ $y_1 \sim \chi^2_{n1}$, และ $y_2 \sim \chi^2_{n2}$ โดยที่ $y_1$ และ $y_2$ เป็นตัวแปรที่เป็นอิสระต่อกัน

2.5.5 ในเรื่องของ statistical inference ยังมีหัวข้อที่สำคัญอีก 2 หัวข้อคือการประมาณค่า (estimation) และการทดสอบสมมุติฐาน(testing hypothesis)

1) Estimation

a. กรณี normal distribution(n > 30)

$P(\text{mean } x - 1.96\sigma < \mu < \text{mean } x + 1.96\sigma) = 0.95$

b. กรณี t-distribution(n < 30)

P(mean x – t(s/ $\sqrt{n}$) < $\mu$ < mean x - t(s/ $\sqrt{n}$) = 0.95

2) Testing hypothesis กรณี population mean มีขั้นตอน(formal step) 3 ขั้นตอน ดังนี้

a. กำหนด

$H_0$ $\mu$ = $\mu_0$

$H_1$ $\mu$ = $\mu_0$

เมื่อ $H_0$ = null hypothesis, $H_1$ = alternative hypothesis, $\mu_0$ = hypothetical value

b. กำหนดระดับความเชื่อมั่น (โดยปรกติทั่วไปจะใช้ = 5%)

c. สุ่มตัวอย่างจากประชากร คำนวณหาค่า mean ของ X (in standard deviation units) ซึ่งก็คือค่า Z (กรณี n > 30) ถ้าค่าของ Z ตกอยู่ใน acceptance region คือยอมรับ $H_0$ ส่วนกรณีอื่นๆ คือปฏิเสธ $H_0$ ยอมรับ $H_1$

**2.6 การหาค่าสัมประสิทธิ์สหสัมพันธ์(correlation coefficient)**

ค่าสัมประสิทธิ์สหสัมพันธ์เป็นค่าทางสถิติที่ใช้วิเคราะห์ความสัมพันธ์ระหว่างตัวแปร 2 ตัวว่ามีความสัมพันธ์กันหรือไม่ มากน้อยเพียงใด และเป็นไปในทิศทางใด ใช้แทนด้วยสัญลักษณ์ $\gamma$

ค่า $\gamma$ มีค่าตั้งแต่ -1 จนถึง +1, ถ้าเท่ากับ | 1 | แสดงว่ามีความสัมพันธ์กันมากที่สุด กรณีเท่ากับ 0 แสดงวาไม่มีความสัมพันธ์กันเลย, ค่า + แสดงความสันพันธ์ไปในทิศทางเดียวกัน ส่วนค่า – แสดงว่าสัมพันธ์ในทางตรงกันข้าม

ความสัมพันธ์ที่เกิดขึ้นนั้น ถ้าไม่มีทฤษฎีทางเศรษฐศาสตร์กำกับไว้ จะไม่สามารถบอกได้ว่า ตัวแปรอะไรเป็นเหตุหรืออะไรเป็นผล แต่ถ้ามีทฤษฎีระบุไว้ เช่นราคาเป็นปัจจัยกำหนดปริมาณ ถ้าสมมุติหาสัมประสิทธิ์สหสัมพันธ์ได้เท่ากับ – 0.5 แสดงว่าราคากำหนดปริมาณได้แต่ไม่มีผลมากนัก

สูตรคำนวณค่า $\gamma$ มีดังนี้

$$\gamma = (n\sum xy - (\sum x)(\sum y))/(\sqrt{n\sum x^2 - (\sum x)^2})(\sqrt{n\sum y^2 - (\sum y)^2})$$

การถอด square root เป็นการถอดทั้งพจน์ในวงเล็บทั้งหมด

ค่าสัมประสิทธิ์สหสัมพันธ์และค่า Chi-square ($\chi^2$) มีความสัมพันธ์กัน ดังนี้

$$\chi^2 = (\gamma^2 N) / (1- \gamma^2)$$

**2.7 Regression analysis**

2.7.1 Ordinary least square(OLS)

Regression analysis ของข้อมูล cross-section โดยวิธี OLS มีวิธีการในทำนองเดียวกันกับข้อมูลอนุกรมเวลา(ดูในเอกสาร macroeconomic model) แต่มีเทคนิคในรายละเอียดและการวัดค่าทางสถิติมากกว่าและต่างกันอยู่บ้าง

ในกรณีของข้อมูลอนุกรมเวลา การสร้างสมการด้วยวิธี regression เพื่อหาค่า parameter มีความหมายว่าเป็นการเอาพฤติกรรมในอดีตมาใช้สำหรับหาหรือทำนายค่าในอนาคต แต่ในกรณีของ cross section data เป็นการตีความว่า เป็นการหาค่า parameter ซึ่งเป็นพฤติกรรมที่ได้จากกลุ่มตัวอย่างที่สำรวจเพื่อเอาไปอธิบายพฤติกรรมโดยรวมหรือของหน่วยอื่นๆ ในประชากรเดียวกันทั้งหมด ณ เวลาเดียวกัน ดังนั้นสมมุติว่าถ้าเราเก็บข้อมูลภาคตัดขวางเพื่อสร้างสมการการบริโภคขึ้นอยู่กับรายได้ในพื้นที่ภาคเหนืออาจไม่สามารถเอามาใช้กับภาคกลางหรือภาคใต้ได้

เนื่องจากข้อมูล cross section ที่นำมาใช้ run linear regression ซึ่งมีลักษณะเป็นราย record เรียงกันไปในแนวของ row(อาจเป็น column ก็ได้แต่ในที่นี้ถือว่า row เป็นกรณีปรกติทั่วไป) ดังนั้นเมื่อนำทุก record มารวมกัน จะได้ชุดข้อมูลมีลักษณะเป็นเสมือน matrix ขนาดใหญ่ ตัวแปรแต่ละตัวคือข้อมูลใน field ต่างๆแต่ละ field เรียงกันตั้งแต่ observationที่ 1 ในแถวที่ 1, observationที่ 2 อยู่ในแถวที่ 2 เรียงกันลงไปจนค่า observation สุดท้าย program SPSS ก็มีการจัดการข้อมูลในลักษณะดังกล่าว ในการ run regression จะต้องกำหนดว่าจะเอาตัวแปรคือค่าใน field ใดเป็นตัวแปรต้น(independent or regressors) และตัวแปรใดเป็นตัวแปรตาม(dependent or regressand) ทั้งในรูปแบบของ single regression และ multiple regression

ถ้าเขียนในรูปของ matrix จะได้เป็น

$$Y = X \beta + \varepsilon$$

เมื่อ Y = matrix ของ dependent variable(regressand) มีขนาด n x1

X = matrix ของ independent variable(regressors) มีขนาด n x k

$\beta$ = matrix ของค่า regression coefficient มีขนาด k x 1

$\varepsilon$ = matrix ของ error term มีขนาด n x 1

n = จำนวน observations

k = จำนวนตัวแปร independent

โดยวิธี ordinary least squares, ค่าของ $\beta$ เท่ากับ

$$\beta = (X' X)^{-1} (X' y)$$

เงื่อนไขคือ

1) $(X' X)$ ต้องสามารถ inverse ได้ หมายความว่า rank[3] ของ X ซึ่งสมมุติให้เท่ากับ k, จะต้องไม่มี column ของ x ใดๆ ที่จะสามารถแสดงค่าได้พอดี(exactly)เท่ากับการ combination (เช่น การบวก ลบ หรือคูณ หาร) ของ k-1 column ที่เหลือ
2) จำนวน observation ต้องมากกว่าจำนวนของ $\beta$(หรือจำนวน column ของ x) นั่นคือ n > k

2.7.2 OLS variance estimator and standard error

Variance ($\sigma^2$)

$$\sigma^2 = (e' e) / n-k$$

Standard error

$$\sigma = \sqrt{\sigma^2}$$

2.7.3 การคำนวณค่า uncentered $R^2$

$$R^2_u = RSS_u/TSS_u = 1- SSE_u/ TSS_u$$

เมื่อ $RSS_u = y' P_x y$

$TSS_u = y' y$

$SSE_u = y' M_x y$

$P_x$ = Projection matrix = $X(X'X)^{-1}X'$

$M_x$ = Annihilator matrix = $I_n - X(X'X)^{-1}X'$

---

[3] ค่า rank หมายถึงจำนวน row(และ column) ที่เป็นอิสระจาก row(และ column) อื่นๆ ในจำนวนที่สูงที่สุด ตัวอย่างเช่น matrix หนึ่ง มีขนาด 5 X 7 โดยที่ค่าใน row ที่ 1 ถึง 5(และ column ที่ 1 ถึง 7) เป็นอิสระต่อกันทั้งหมด แสดงว่า matrix นี้มีค่าเท่ากับ 7 ซึ่งมีจำนวนสูงสุด แต่ถ้าสมมุติว่า column ที่ 7 ได้จาก column ที่ 3 คูณกับ column ที่ 5(หรืออื่นๆ เช่น column 3 คูณกับค่าคงที่ค่าหนึ่ง) แสดงว่า column ที่ 7 ไม่เป็นอิสระ ไป 1 column ดังนั้น matrix นี้จึงมี rank เท่ากับ 6

2.7.4 R-bar square

$$(\text{R-bar})^2 = SSE_{n-1} / TSS_{n-k}$$

ค่า R-bar square อิสระจากจำนวน regressor ดังนั้นการเปรียบเทียบระหว่าง 2 model(ซึ่งแต่ละ model อาจมีจำนวน regressor ไม่เท่ากัน) จึงต้องใช้ R-bar square เสมอ

2.7.5 สมมุติฐานของ regressors

การ run linear regression ของ cross sectional data จำเป็นต้องมีสมมุติฐานกำหนดให้กับค่า regressor ทั้งนี้ขึ้นอยู่กับจำนวนข้อมูลที่มี(observation)ว่าเป็นข้อมูลขนาดเล็กหรือขนาดใหญ่

จำนวนข้อมูลที่เหมาะสมเพียงพอกับการ run regression ของ cross sectional data ขึ้นอยู่กับการรู้ลักษณะของค่า error term ถ้า error term เป็นอิสระต่อกันและมีการกระจายที่ด้านซ้ายและขวามีรูปร่างเหมือนกัน(identically distribution or asymptotic distribution) เช่นเป็นรูประฆังคว่ำ ในลักษณะนี้ จำนวน observation ประมาณ 30 ค่า(finite or small sample) ก็อาจจะเพียงพอ แต่ถ้าไม่เป็นอิสระต่อกันหรือการกระจายไม่ indentically ซึ่งกันและกัน ในกรณีนี้อาจจำเป็นต้องใช้ตัวอย่างจำนวนมากถึง 100 ถึง 1,000 ตัวอย่าง(large sample) หรืออาจจำเป็นต้องมีจำนวนมากกว่านี้

ในกรณีของ cross sectional analysis นั้น การกำหนดสมมุติฐานให้กับ regressor และการทดสอบสมมุติฐานค่าต่างๆ(การทดสอบสมมุติฐานก็คือการวัดว่าผลที่ได้จากข้อมูลที่รวบรวมได้จาก observation ต่างๆมีความเป็นจริงที่ตรงกับทฤษฎี) ขึ้นอยู่กับขนาดของจำนวน observation ว่าเป็น small scale sample(finite) หรือเป็น large scale sample(non-finite) อย่างไรก็ตาม โดยทั่วไปแล้ว จะมีสมมุติฐานประกอบด้วย ดังนี้

1) Linearity หมายถึงว่า regressor อยู่ร่วมกับ parameter ในลักษณะ multiplicative และ error term อยู่ในรูป additive
2) Conditional mean หมายถึงว่า error term มีค่า mean เท่ากับศูนย์ ไม่ว่า x จะเป็นเท่าไร
3) rank ของ x มีค่าเท่ากับ k และมีค่า propabilty เท่ากับ 1, โดยที่จำนวนของ observation, n ต้องมากว่า k, หรือจำนวน regressor เพื่อให้ matrix ของ $X'X$ สามารถที่จะ invert ได้, ส่วน propability เท่ากับ 1 นั้นเพื่อให้มั่นใจว่า x แต่ละตัวไม่ correlate กัน(ทำนองเดียวกันกับ multicollinearity ในข้อมูล time series)
4) Conditional homoskedasticity หมายความว่า ค่า residual มี variance ที่เหมือนกัน
5) Conditional correlation หมายความว่า error term แต่ละค่าต้องไม่สัมพันธ์กัน(ทำนองเดียวกับ autocorrelation ในข้อมูล time series)

6) Conditional normality หมายถึงการกระจายตัวของค่า error term เป็น normal

2.7.6 Hypothesis testing

Regression ของ cross sectional data มีหลักการของการทดสอบสมมุติฐาน ในทำนองเดียวกับข้อมูล time series แต่วิธีการสร้าง form ของการทดสอบจะเป็นในรูปของ matrix ค่าสถิติที่ใช้ทดสอบที่สำคัญประกอบด้วย

1) t-test เป็นการทดสอบสมมุติฐานว่าค่าของ parameter ไม่เท่ากับ 0
2) Wald test เป็นการทดสอบระยะห่างระหว่าง $R\beta$ และ r, ถ้า $H_0$ เป็นความจริง(ยอมรับ $H_0$) หมายความว่า $R\beta - r \approx 0$, R คือ functional form ของ $R^k$ to $R^m$ และ r เป็น vector ขนาด m x 1
3) Likelihood ratio test เป็นการทดสอบ relative probability ของ observed data
4) Lagrange multiplier test เป็นการทดสอบ
   $\min_\beta (y-X\beta)' (y-X\beta)$ , subject to $R\beta - r = 0$

**3. การใช้งาน program SPSS**

Program SPSS(Statistical Package for the Social Science for Windows) เป็นโปรแกรมที่สามารถใช้ประโยชน์ได้ทั้งการจัดการข้อมูลและการวิเคราะห์ข้อมูลประเภท cross section data จุดมุ่งหมายของการนำเสนอหัวข้อการใช้งาน program SPSS ในที่นี้เพื่อเป็นการให้คำแนะนำในเบื้องต้นของการใช้ program ทางสถิติสำหรับ cross sectional analysis ตามที่ได้อธิบายในหัวข้อที่ 2 (หมายเหตุ ผู้เขียน(ผู้รับจ้าง) ได้เคยศึกษาและได้ใช้งาน program SPSS กับข้อมูลผลการสำรวจภาวะเศรษฐกิจและสังคมของครัวเรือน (observation > 100,000 records) เมื่อคราวไปปฏิบัติงานเป็นที่ปรึกษาที่ สวค. แต่ก็เป็นเพียงการใช้งานจริงในเบื้องต้น เฉพาะการจัดการข้อมูลที่ไม่ซับซ้อนและ run regression อย่างง่ายของแบบจำลองการบริโภคของครัวเรือนที่ขึ้นกับรายได้รวมกับตัวแปรอื่น พบว่าเป็น program ที่ไม่ยุ่งยากมากนัก ถ้า ขว. จะต้องทำงานวิจัยต่อไปในอนาคตเห็นว่า ควรมีการจัดหา program ดังกล่าว รวมทั้งการส่งบุคลากรไปฝึกอบรมเพื่อใช้งานโปรแกรมดังกล่าว)

มีเอกสารพิมพ์เผยแพร่คำอธิบายเกี่ยวกับการใช้งาน program SPSS อยู่หลายเล่ม ซึ่งเล่มที่ผู้เขียนใช้คือ “การวิจัยและวิเคราะห์ข้อมูลทางสถิติด้วย SPSS และ AMOS” เขียนโดยรศ. ธานินทร์ ศิลป์จารุ(พ.ศ. 2555) ซึ่งเป็นการอธิบายบน version 15

เมื่อเริ่มเรียกใช้งาน program SPSS จะมีหน้าต่างของการทำงานที่คุ้นเคยตามลักษณะคล้ายกับ MS Excel

ส่วนประกอบของโปรแกรม SPSS ประกอบด้วย

1) Data editor windows
2) Syntax editor window

3) Output window

4) Draft output window

5) Script window

**3.1 การนำเข้าข้อมูล**

การนำเข้าข้อมูลสามารถทำได้ทั้งการ key ข้อมูลเข้าไปโดยตรงและ การ import จาก Excel (หรืออื่นๆ เช่น notepad หน้าต่างที่ใช้สำหรับการนำเข้าข้อมูลคือ Data editor windows

3.1.1 กรณีนำเข้าข้อมูลโดยตรง

1) ตั้งค่าตัวแปร ในแถบคำสั่ง variable view จะมีหัวตารางประกอบด้วย name, type, width, decimals, label…….ให้กำหนดค่าต่างๆ ลงไป
2) เลือกเซลล์ แถวและคอลัมน์
3) ป้อนข้อมูล
4) สามารถทำการคัดลอกหรือย้ายข้อมูลตามที่ต้องการได้
5) บันทึก file ข้อมูล (นามสกุลเป็น xxx.sav)

3.1.2 กรณีนำเข้าจาก MS-Excel

1) ถ้ามีข้อมูลอยู่แล้วต้องจัด form ของข้อมูลให้ตรงกับที่ SPSS จะเรียก หรือถ้า key ข้อมูลใหม่ต้องเริ่มจากการ,เปิดโปรแกรม Excel, ใส่ชื่อตัวแปรในแถวแรกบนหัวของคอลัมน์ทุกคอลัมน์ ตามชื่อตัวแปรต่างๆ
2) Key ข้อมูลแต่ละตัวเรียงกันลงมาตามแนวคอลัมน์ เช่น สมมุติคอลัมน์ที่ 1 ชื่อตัวแปรรายได้ของธุรกิจ ให้ใส่ข้อมูลรายได้ของ observation ที่ 1 ในแถว 2 (แถว 1 เป็นชื่อตัวแปร), observation ที่ 2, 3, 4…..ในแถวต่อๆ มาเรียงกันลงมา หรือในกรณีถ้ามีข้อมูลอยู่แล้วก็ต้องจัด form ของ sheet ให้เป็นตามแนวทางดังกล่าวดังกล่าว อาจใช้การ insert แถวบนสุดแล้วใส่ชื่อตัวแปรไปก็ได้ถ้าของเดิมไม่มีไว้
3) Save file excel ที่ key ข้อมูลตาม form หรือจัดการตาม form ดังกล่าวแล้ว
4) เปิด SPSS, หน้าต่าง data editor, เลือกคำสั่ง read text data, กำหนดประเภท file ที่จะเรียกมาเป็น MS-Excel
5) Open file name และ file type ที่ต้องการ, กำหนดชื่อตัวแปรแถวแรก, กำหนด sheet, กำหนดช่วง cell แล้ว กด OK เพื่อ import ข้อมูลเข้ามา
   <u>หมายเหตุ</u> คำอธิบายที่เขียนนี้เป็นการอธิบายหลักโดยคร่าวๆ เท่านั้น ในการทำจริงจะมีรายละเอียดมากว่านี้ ผู้ใช้จะต้องดูจาก window และแถบต่างๆ รวมทั้งดูวิธีใช้จากแนวทางที่ระบุในคู่มือประกอบกันไป

**3.2 การจัดการข้อมูล**

สามารถดำเนินการได้หลายลักษณะ ที่สำคัญคือ

3.2.1 การสร้างตัวแปรใหม่จากการคำนวณและ function ตามที่ SPSS กำหนดไว้, การสร้างใช้แถบคำสั่ง transform, compute variable

3.2.2 การสร้างตัวแปรใหม่แบบ if case

3.2.3 การเลือกตัวแปรย่อยโดยวิธี select cases

**3.3 การวิเคราะห์ค่าทางสถิติ**

3.3.1 การวิเคราะห์ค่าทางสถิติทั่วไป

การวิเคราะห์ค่าสถิติทั่วไปยังคงใช้หน้าต่าง data editor, ใช้แถบคำสั่ง analyze, descriptive statistic หลังจากนั้นก็เลือกใช้คำสั่งย่อยตามที่ต้องการ เช่น frequencies

1) การหาค่าสถิติของตัวแปรเดียว เช่น ค่าสูงสุด ค่าต่ำสุด ค่าความถี่ ค่าตัวกลาง mean, median, mode ค่าการเบี่ยงเบน(standard deviation)

นอกเหนือจากที่กล่าวแล้ว ภายใต้คำสั่ง descriptive statistic สามารถที่จะเลือกคำสั่งย่อยคือ crosstabs ก็ได้ซึ่งจะเป็นประโยชน์ในกรณีที่ต้องการวิเคราะห์แบบแจกแจงค่าความถี่ 2 ทาง

ผลที่ได้สามารถกำหนดให้ออกรายงานเป็นรูปกราฟก้ได้เช่นกัน

2) การเปรียบเทียบระหว่างตัวแปร เช่น compare means, หาค่า correlation, คำนวรหค่า Chi-square, t-test, ANOVA เป็นต้น

3.3.2 การ run regression

เป็นการใช้คำสั่งภายใต้หน้าต่าง data editor เช่นกัน ไปที่แถบคำสั่ง analyze, เลือกคำสั่ง regression, linear regression หลังจากนั้นเป็นการเลือกตัวแปรที่จะกำหนดเป็น dependent variable และ independent variable และกดปุ่ม OK เพื่อ run linear regression ที่ต้องการ

โปรแกรม SPSS จะให้ค่าสิติต่างๆ ทั้งค่า estimated parameter, ค่า standard error, $R^2$, adjusted-$R^2$, ค่า t-statistic, ค่า P(เพื่อใช้ทดสอบว่า parameter ดังกล่าว significant หรือไม่(ถ้าค่าที่คำนวณ > $\alpha$ .05 แสดงว่าไม่ significant)

การจดจำค่าต่างๆ ดังกล่าวนี้ค่อนข้างยุ่งยากเพราะมีรายละเอียดมาก จำเป็นต้องฝึกปฏิบัติจากข้อมูลจริง และเรียนรู้ไปตามคู่มือทีกำหนดจะเข้าใจได้ไม่ยากนัก(แต่ที่สำคัญคือต้องทำความเข้าใจความหมายและการตีความของค่าต่างๆที่ได้ และใช้ประโยชน์การวิเคาะห์จากค่าดังกล่าว)

---