Regression Analysis

ท่านผู้อ่านหลายท่านคงจะเคยได้อ่านคำแปลภาษาไทยของคำนี้ว่า "การวิเคราะห์การถดถอย" ถึงแม้คำแปลนี้จะเป็นที่รับรู้ในภาษาไทยมานานแล้ว แต่โดยส่วนตัวแล้วผมคิดว่าเป็นคำแปลที่ทำให้ผู้อ่านรู้สึกหดหู่พิกล แต่ผู้เขียนก็ไม่ได้บอกว่าคำที่ใช้ในปัจจุบันไม่ถูกต้อง และที่ถูกต้องควรจะใช้คำว่าอะไร เพียงแต่อยากให้ท่านผู้อ่านลองอ่านเนื้อหาเบื้องต้นดูก่อน แล้วจะทราบว่า คำว่า Regess นั้นที่ถูกแล้วภาษาไทยเราควรจะใช้คำว่า " ถดถอย" หรือไม่

ท่านผู้อ่านเคยมีคำถามในใจคล้ายๆกับคำถามต่อไปนี้บ้างหรือไม่

1. เปิดเครื่องปรับอากาศ วันละ 5 ชั่วโมง เดือนนี้จ่ายค่าไฟ 560 บาท ถ้าอยากจ่ายเดือนละ 450 บาท หรือน้อยกว่าควรจะเปิดวันละกี่ชั่วโมง (อุปกรณ์อย่างอื่นใช้เหมือนปกติ)

2. ถ้าขับรถด้วยความเร็วเฉลี่ย 90 กิโลเมตรต่อชั่วโมง เติมน้ำมันเต็มถังจะวิ่งรถได้รวม 560 กิโลเมตร ถ้าขับที่ความเร็วเฉลี่ย 110 กิโลเมตรต่อชั่วโมง จะวิ่งได้กี่ กิโลเมตร แล้วถ้าเป็น 130 กิโลเมตรต่อชั่วโมง จะได้กี่ กิโลเมตร ถ้าเติมน้ำมันเต็มถังเหมือนกัน

3. ถ้าตั้งอุณหภูมิหม้อต้มน้ำที่ 100 องศาเซลเซียส จะสามารถฆ่าเชื้อแบคทีเรียได้ ร้อยละ 85 ภายใน 1 นาที อยากทราบว่าถ้าต้องการให้ฆ่าได้ ร้อยละ 90 ในเวลาเท่าเดิม จะต้องใช้อุณหภูมิเท่าใด และถ้าต้องการให้ฆ่าได้ หมด ร้อยละร้อย จะต้องใช้อุณหภูมิ เท่าใด

โจทย์ตัวอย่างทั้งสามข้อนั้น กำลังชี้ให้ท่านเห็นถึงความสัมพันธ์ ของตัวแปรสองตัว หากให้ท่านเลือกว่าจะใช้เครื่องมือทางคณิตศาสตร์อะไรในการแก้ปัญหาโจทย์ลักษณะนี้ ผมเชื่อว่าหลายท่านคงกำลังนึกถึง บัญญัติไตรยางศ์ ที่เราเคยเรียนมาตั้งแต่สมัยชั้นประถม ซึ่งผมกำลังจะบอกว่าท่านนึกถึงเฉยๆนะได้ แต่มาใช้กับโจทย์ในลักษณะนี้ไม่ได้นะครับ ผมมีเหตุผลที่จะบอกท่านอย่างนี้ครับ

1. บัญญัติไตรยางศ์ จะใช้ได้ก็ต่อเมื่อตัวแปรทั้งสอง เป็นตัวแปรที่ค่า ไม่มีความคลาดเคลื่อน เช่น ขนมกล่องนี้ราคา 10 บาท ถ้าซื้อ 10 กล่อง ต้องจ่าย 100 บาท ตัวแปรที่ว่าคือ ราคาต่อกล่องก็คงที่ คือ 10 บาท ตัวแปรอีกตัวคือจำนวนที่จะซื้อ คือ 10 กล่อง ก็ไม่มีความคลาดเคลื่อนเลย

2. โจทย์หรือคำถามสามข้อที่ผมกล่าวถึง ที่ผ่านมานั้น ท่านลองสังเกตดีๆ จะเห็นว่า ถึงจะเป็นโจทย์ที่กล่าวถึงความสัมพันธ์กันของสองตัวแปร แต่จะมีเพียงหนึ่งตัวแปรเท่านั้นที่จะไม่มีความคลาดเคลื่อนเลย แต่อีกตัวแปรที่เหลือนั้น โดยธรรมชาติของมันจะมีค่าคลาดเคลื่อนตลอดเวลา ไม่มากก็น้อย เช่น ตั้งอุณหภูมิ 100 องศา แล้วฆ่าเชื้อแบคทีเรียได้ ร้อยละ 85 ถ้าทำการทดลองอีกที ภายใต้สถานการณ์เหมือนเดิมแท้ๆ อาจจะฆ่าได้ถึงร้อยละ 87 ทำซ้ำอีกรอบเหมือนเดิม อาจจะฆ่าได้เพียงร้อยละ 83 ไม่ใช่ทดลองไม่ดี แต่นี่เป็นความคลาดเคลื่อนที่เป็นธรรมชาติ เป็นเหตุสุดวิสัยที่จะไปควบคุมให้ได้ค่าเดิมทุกครั้งได้ (ถ้าใครทำได้แปลว่าท่านต้องทำอะไรสักอย่างไม่ดีแน่ๆ)

ด้วยเหตุผล 2 ข้อดังกล่าว การวิเคราะห์ถึงความสัมพันธ์ของตัวแปรสองฝั่งจึงต้องใช้เทคนิคทางคณิตศาสตร์ที่พิเศษกว่าบัญญัติไตรยางศ์ ซึ่งก็คือ " Regression Analysis " และแทนที่เราจะเรียกว่าการวิเคราะห์หา

ความสัมพันธ์ เราก็จะเรียกว่า การประมาณการ (Prediction) แทน เมื่อเป็นเช่นนี้ ตัวแปรฝั่งที่ไม่มีค่าคลาดเคลื่อน เราจะเรียกว่าตัวประมาณการ (Predictor) โดยใช้ สัญลักษณ์แทนคือ X ตัวแปรที่มีความคลาดเคลื่อน เราก็จะเรียกว่า ตัวตอบสนอง (Response) สัญลักษณ์แทนคือ Y โดยที่

$$Y = F(X)$$

ผลการวิเคราะห์ที่ได้ เราจะได้สมการหรือฟังก์ชันคณิตศาสตร์ที่แสดงถึงความสัมพันธ์กันของทั้งสองตัวแปร เช่น

$$\% \text{ Bacteria killed } (Y) = 67.45 + 0.214 * \text{ Temperature}(X)$$

ความสัมพันธ์ที่เขียนแทนด้วยฟังก์ชันคณิตศาสตร์ดังกล่าวเราจะเรียกว่า Model หรือ Mathematical Model และฟังก์ชันคณิตศาสตร์ที่ได้ จะสามารถนำไปประมาณการ ตัวแปรฝั่งที่มีค่าคลาดเคลื่อนได้ โดยใช้ค่าของตัวแปรฝั่งที่มีค่าไม่คลาดเคลื่อน แปลว่าเมื่อเรารู้ค่าตัวแปรที่ค่าไม่คลาดเคลื่อน และรู้ฟังก์ชั่นคณิตศาสตร์แสดงความสัมพันธ์ แล้วเราก็สามารถรู้ค่าตัวแปรฝั่งที่มีค่าคลาดเคลื่อนได้ Mathematical Model ดังกล่าวจึงเรียกว่า Transfer function ในที่สุด มาถึงตรงนี้ท่านผู้อ่านอย่าสับสนนะครับ โปรดจำไว้ว่า ไม่จำเป็นที่ Mathematical Model ทุกอันต้องเป็น Transfer function เฉพาะที่เกี่ยวข้องกับสองตัวแปรหรือมากกว่าขึ้นไปที่สามารถอธิบายความสัมพันธ์ในลักษณะใช้ฝั่งหนึ่งประมาณการ อีกฝั่งหนึ่งได้เท่านั้นจึงจะเรียกว่า Transfer function

ในการศึกษา ถึงความสัมพันธ์ของตัวแปรสองฝั่งโดยใช้ Regression analysis นั้นสามารถที่จะใช้ได้กับหลายลักษณะความสัมพันธ์ และปริมาณตัวแปร เช่น

1. Simple linear regression analysis : ชื่อก็บ่งบอกว่าใช้ได้เมื่อใด ก็คือจะใช้เมื่อเราต้องการวิเคราะห์ความสัมพันธ์ระหว่าง สองตัวแปร และความสัมพันธ์ระหว่างสองตัวแปรดังกล่าวจะต้องเป็นในลักษณะเชิงเส้น ดังตัวอย่างง่ายๆ

$$\% \text{ Bacteria killed } (Y) = 67.45 + 0.214 * \text{ Temperature}(X)$$

2. Multiple linear regression analysis : จะใช้เมื่อเราต้องการวิเคราะห์ความสัมพันธ์ เมื่อมีตัวแปรที่เป็น Predictor มากกว่า 1 ตัวขึ้นไป แต่ความสัมพันธ์ของตัวแปรทั้งสองฝั่ง ยังคงเป็นแบบเชิงเส้นตรง ยกตัวอย่างในกรณีคำถามที่ 3 ที่ผ่านมานั้น นอกจากอุณหภูมิจะมีผลต่อจำนวนเชื้อแบคทีเรียที่ถูกฆ่าแล้ว เวลาที่แช่ภาชนะเพาะเชื้อแบคทีเรียในน้ำรอน ก็เป็นตัวแปร ที่มีผลด้วยเช่นกัน เพราะถ้าอุณหภูมิเท่าเดิม การใช้เวลา 1 นาที กับ 2 นาทีสามารถฆ่าเชื้อแบคที่เรียได้จำนวนแตกต่างกัน ด้วย สมมติว่าความสัมพันธ์เป็นเชิงเส้นตรง ตัวอย่าง Mathemical model จะเป็นดังนี้

$$\% \text{ Bacteria killed } (Y) = 36.415 + 0.412 * \text{ Temperature}(X_1) + 4.85 * \text{Time}(X_2)$$

3. Polynomial regrssion analysis : จะใช้เมื่อเราต้องการวิเคราะห์ถึงความสัมพันธ์ที่ไม่เป็นเชิงเส้นตรง รวมถึงกรณีมีตัวแปร Predictor มากกว่า 1 ด้วย ยกตัวอย่างเช่น ถ้ากรณีเปิดเครื่องปรับอากาศนั้น นอกจากจำนวนชั่วโมงที่เปิดจะมีผลต่อจำนวนหน่วยไฟฟ้าที่ใช้แล้ว อุณหภูมินอกห้อง ก็ส่งผลด้วยเหมือนกันและไม่

เป็นเส้นตรงด้วย การวิเคราะห์ก็จะยิ่งซับซ้อน และยุ่งยากมากขึ้นไปอีก Mathemical Model จะเป็นดังตัวอย่างต่อไปนี้

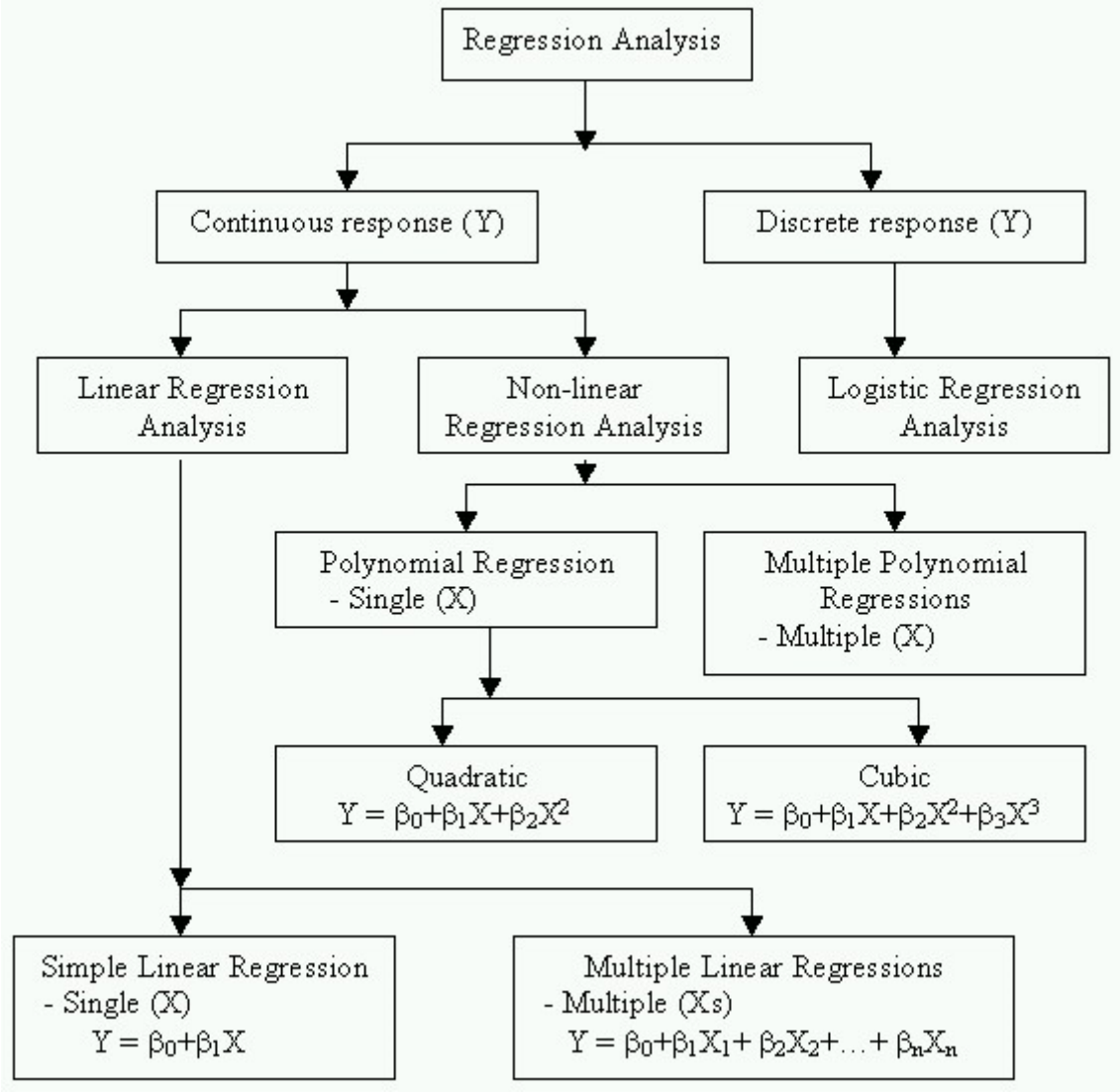
$$\text{Power consumtion (Unit)} = 1.542 + 0.859 * \text{Time} + 0.587*(\text{External Temperature})^2$$

ในเมื่อลักษณะความสัมพันธ์ของตัวแปร มีทั้งแบบเชิงเส้นและแบบไม่เชิงเส้น (Polynomial) หลายท่านอาจจะสงสัยว่า เราจะรู้ได้อย่างไรว่าความสัมพันธ์ของตัวแปรเป็นเชิงเส้นหรือ ไม่เป็นเชิงเส้น คำตอบคือท่านจะต้องใช้การทำ Scatter plot เข้าช่วยอย่างหลีกเลี่ยงไม่ได้ ดังนั้นการวิเคราะห์ความสัมพันธ์โดยใช้ Regression analysis ท่านจำเป็นต้องทำกราฟแสดงจุดตัดของตัวแปรทั้งสองฝั่งให้ได้ และ Scatter plot คือเทคนิคที่ช่วยท่านได้ดีที่เดียว ให้ท่านผู้อ่านย้อนกลับไปอ่านการทำ Scatter plot โดยใช้ MS Excel เพื่อความเข้าใจ

4. Logistic regression analysis กรณีที่ Y มีค่าเพียงสองสถานะ เช่น No ,Yes เป็นต้น แต่ X เป็นค่าแบบต่อเนื่องปกติ

ภาพโดยรวมของการวิเคราะห์ข้อมูลโดย Regression Analysis

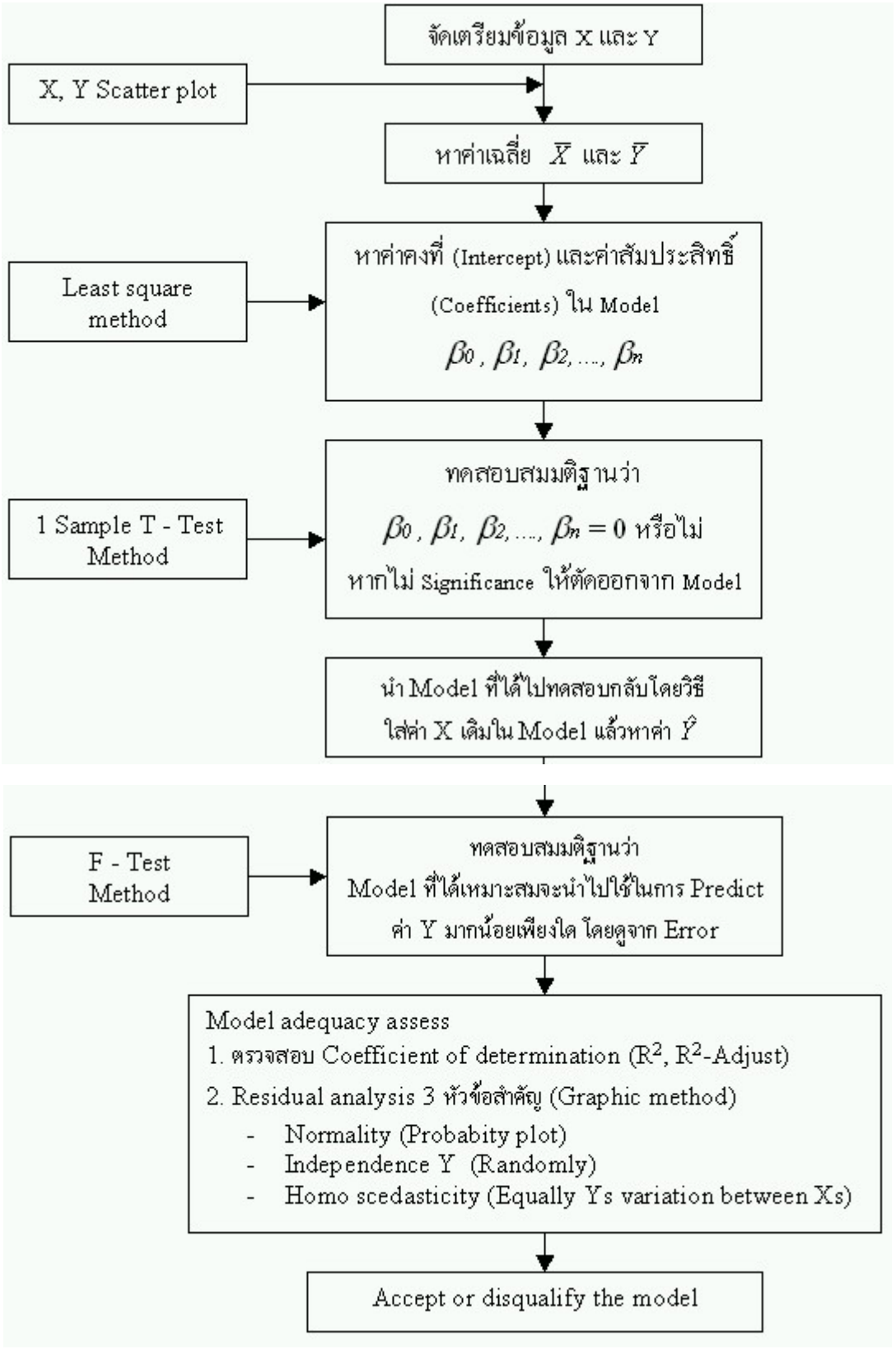
Regression Analysis ถือเป็นเครื่องมือทางสถิติที่มีการประยุกต์ใช้ในการประมวลผลข้อมูลในงานวิจัยค่อนข้างมาก นักสถิติยุคก่อนได้คิดค้นทฤษฎีเกี่ยวกับ Regression เอาไว้มากมาย ภาพต่อไปนี้แสดงให้เห็นว่าท่านควรจะเลือกใช้เครื่องมือวิเคราะห์แบบใด ถึงจะเหมาะและตรงกับข้อมูลที่ท่านมีอยู่



ภาพ แสดงการจัดแบ่ง Regression Analysis ตามชนิดของข้อมูลที่ต้องการจะวิเคราะห์

ลำดับขั้นตอนของการวิเคราะห์ข้อมูลโดย Regression Analysis

เมื่อท่านตัดสินใจแล้วว่า จะต้องใช้ Regression Analysis ในการวิเคราะห์ข้อมูลของท่านแล้ว ขั้นตอนและลำดับการวิเคราะห์จะเป็นดังแผนภาพต่อไปนี้ (เหมาะกับ Simple linear regression และ Multiple linear regression ที่ผู้เขียนคิดว่า ผู้อ่านหรือผู้ที่จะทำการวิเคราะห์ข้อมูล จะเลือกใช้มากกว่า วิธีอื่นๆ)



ภาพ แสดงลำดับการวิเคราะห์ข้อมูลโดยใช้ Regression Analysis

การทดสอบนัยสำคัญทางสถิติ (Statistical significance)

1. T-Statisitic เมื่อเราได้ Model มาแล้วจะต้องพิสูจน์ทางสถิติ ค่าคงที่ (b_0)และสัมประสิทธิ์ของตัวแปรอิสระทุกค่า ($b_1, b_2, \ldots, b_n$) ว่ามีนัยสำคัญต่อ Model หรือไม่ โดยตั้งสมมติฐาน

$$H0 : \beta_i = 0$$
$$Ha : \beta_i \neq 0$$

ถ้าเราต้อง Accept H0 โดยดูจากค่า T ที่คำนวณได้ถ้าน้อยกว่าค่า T-critical ก็แปลว่า ค่าคงที่หรือสัมประสิทธิ์ของตัวแปรอิสระตัวนั้นๆไม่มีนัยสำคัญต่อ Model ก็ตัดออกได้ โดยจะไม่ทำให้ Model นั้นเกิดความแตกต่างแต่อย่างใด ในทางตรงกันข้ามเราต้อง Reject H0 เมื่อ T ที่คำนวณได้มากกว่า T-critical ก็แปลว่าค่าคงที่หรือสัมประสิทธิ์ของตัวแปรอิสระตัวนั้นๆมีนัยสำคัญต่อ Model ไม่สามารถตัดออกได้ นี่เองที่ทำให้ Regression analysis ต้องมี T-test หรือมีค่า t ปรากฏในตารางผลการวิเคราะห์

2. F-Statistic สุดท้ายเราได้ Model ที่ถือว่าดีที่สุดเท่าที่จะหาได้ แต่ก็ใช่ว่า Model ดังกล่าวจะใช้ในการ Predict ค่า Y ได้ถูกต้อง เราจำเป็นต้องพิสูจน์ทราบว่า Model ที่ได้นั้นเมื่อนำไป Predict ค่า Y แล้วจะมีความคลาดเคลื่อนมากแค่ไหน หรือก็คือ มี Error ระหว่าง Y และ $\hat{Y}$ มากแค่ไหน

H0: Error ที่เกิดขึ้นที่ (Y) เกือบทั้งหมดมาจากตัวแปรอิสระ

Ha: Error ที่เกิดขึ้นที่ (Y) ส่วนน้อยเท่านั้นที่มาจากตัวแปรอิสระ

หากผลการทดสอบด้วย F-Test พบว่า Accept Ho หรือ F-Statistic ที่ได้มีค่าต่ำกว่าค่า F-critical ก็ให้ถือว่า Model นั้นมีความผิดพลาดสูง จนไม่อาจยอมรับให้นำไปใช้ต่อไปได้ ก็ถือว่าอย่างอื่นก็ไม่ต้องวิเคราะห์ต่อ ในทางตรงกันข้ามหากผลการทดสอบด้วย F-Test พบว่า Reject Ho หรือ F-Statistic ที่ได้มีค่ามากกว่าค่า F-critical ก็ให้ถือว่า Model นั้น เมื่อนำไป Predict ค่า Y แล้วมีความผิดพลาดน้อยสามารถยอมรับได้ จึงเป็นเหตุให้ Regression analysis ต้องมี ANOVA เพราะใน ANOVA มี F-Test อยู่นั่นเอง

3. Coefficient of determination ใช้พิสูจน์ว่า Model ที่ได้นั้นมีที่มาที่ดีพอจะใช้ Model จากผลการวิเคราะห์ไป Prdedict ค่า Y ในอนาคตได้หรือไม่ แม้ว่า F-Test จะบอกว่า Model มีความผิดพลาดต่ำแค่ไหนก็ตาม แต่หากที่มาของการเก็บข้อมูลก่อนการวิเคราะห์ ไม่เหมาะสม ก็ยังถือว่า Model นั้นที่ดีได้อาจเพราะเหตุบังเอิญ

- R^2 เป็นค่าที่บ่งบอกว่าข้อมูลดิบของการวิเคราะห์นั้น เหมาะสมหรือไม่ จะมีค่าระหว่าง 0 - 1 ยิ่งเข้าใกล้ 1 ก็ยิ่งดี โดยทั่วไปควรมีค่า 0.6 ขึ้นไป แต่ก็ไม่ได้มีกฎเกณฑ์แน่นอนตายตัว

- R^2-Adjust เป็นค่าที่บ่งบอกว่า R^2 ที่ได้นั้นเหมาะสมจริงไหม โดยจะทำการลด Sample (N) ลง 1 ตัว แล้วหาค่า R^2 ใหม่อีกครั้ง เลยเรียกว่า Adjust หากมีค่าต่ำกว่า R^2 มากผิดปกติ ก็ให้สรุปว่า Sample size ต่ำเกินไป หรือ R^2 มี Sensitivity ต่อการเปลี่ยนแปลง N มากเกินไป มีโอกาสที่ Model จะผิดพลาดก็สูงทีเดียว ที่เหมาะสม ค่า R^2-Adjust จะต้องต่ำกว่า R^2 เพียงเล็กน้อยเท่านั้น จึงจะถือว่าการทดลองครั้งนี้ เก็บข้อมูลมาดี Sample size เหมาะสม หากท่านกำลังใช้ Regression analysis แล้ว ค่า R^2 และ R^2-Adjust นี้สามารถจะทำให้ Model ที่ได้นั้น ไม่อาจจะยอมรับให้ใช้ได้ ต้องกลับไปดำเนินการเก็บข้อมูลเพิ่มและเริ่มทำการวิเคราะห์ใหม่อีกครั้งหนึ่ง แม้ว่า F-test จะปรากฏผลว่า Model นั้นมีค่าความคลาดเคลื่อนต่ำก็ตาม

ท่านผู้อ่านอาจจะสงสัยว่าทำไมต้อง Proof หลายชั้นเหลือเกิน จริงครับ การวิเคราะห์ไม่ใช่เรื่องยากแต่การจะ

Proof ว่าผลการวิเคราะห์นั้น ใช้ได้ไหม นี่เป็นเรื่องยากกว่าเยอะ

เนื้อหาเฉพาะของ Regression analysis จะนำเสนอในหัวข้อต่อไป

[HOME] [CONTENTS]

หลักการพื้นฐานของ Simple Linear Regression Analysis

ชื่อก็บอกอยู่แล้วว่าง่ายและไม่ซับซ้อนที่สุดในส่วนของ Regression analysis โดยแท้จริงแล้วก็คือ Regression ที่มี ตัวแปรที่เรารู้ค่า (Predictor) และตัวแปรที่เราไม่รู้ค่า (Response) อย่างละ 1 ตัวเท่านั้น ผู้เขียนขอเริ่มเข้าสู่เนื้อหาโดยการเริ่มวิเคราะห์ตัวอย่างให้เห็นภาพ และพื้นฐาน ที่ไปที่มา ของ Regression ก่อน ผู้เขียนเชื่อว่าตำราหลายๆตำราเกี่ยวกับเรื่อง Regression นี้ จะเริ่มด้วยการหาสมการ ซึ่งผู้เขียนเห็นว่า ข้อเสียของวิธีนี้คือผู้อ่านจะไม่เข้าใจว่าทำไมถึงได้สมการแบบนั้น มันมีขั้นตอนหรือหลักการคิดมาได้อย่างไร

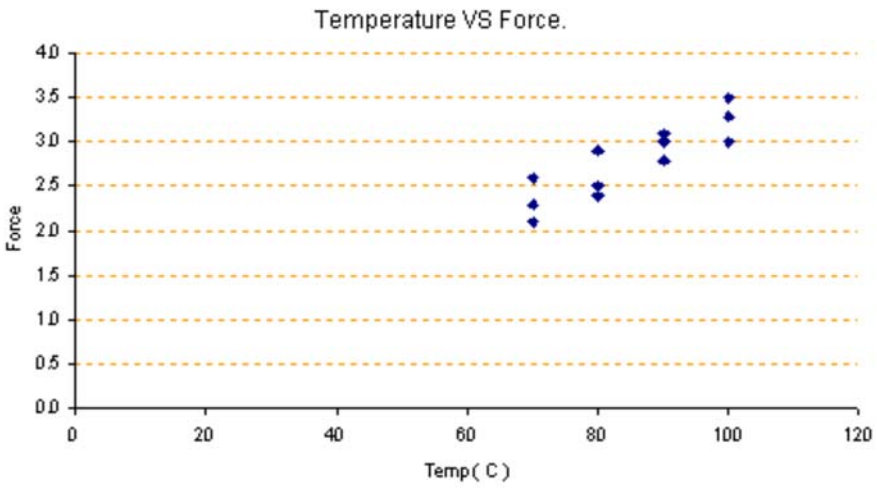
ก่อนอื่นเราต้องทำความเข้าใจก่อนว่า จุดประสงค์ของการใช้ Regression Analysis ก็เพื่อต้องการหาสมการความสัมพันธ์ (Transfer function) ของตัวแปรฝั่งที่เรารู้ค่า (Predictor) กับฝั่งที่เราไม่รู้ค่า (Response) เพื่อที่จะนำไปสู่การคาดการณ์หรือประมาณค่า ของตัวแปรที่เราไม่รู้ค่าได้ในที่สุด และที่สำคัญการจะนำสมการความสัมพันธ์ไปใช้ได้ จะต้องมีการตรวจสอบ (Verify) เสียก่อนว่าสมการที่ได้มานั้นมีความถูกต้อง พอที่จะใช้เป็นสมการในการคาดการตัวแปรที่ไม่รู้ค่าได้จริงหรือไม่

ตัวอย่าง ในการศึกษาเรื่องความสามารถทนแรงดึงของกาว Epoxy ที่จะใช้ในการยึดชิ้นงาน 2 ชิ้นเข้าด้วยกัน โดยขั้นตอนคือ เมื่อหยอดส่วนผสมกาวลงบนชิ้นงาน A แล้วนำชิ้นงาน B มาติดเข้า แล้วต้องเอาเข้าอบด้วยความร้อน เพื่อให้กาวแห้งและชิ้นงาน A และ B ติดกันตามต้องการ ผู้ศึกษาต้อการทราบความสัมพันธ์ระหว่าง อุณหภูมิที่ใช้ในการอบกับความสามารถในการทนแรงดึงของกาวหลังอบ มีแตกต่างกันอย่างไร โดยได้ทำการทดลอง 3 ตัวอย่างต่อการทดลอง 1 รอบ และแต่ละรอบจะตั้งอุณหภูมิไว้คงที่ ที่ค่าที่ต้องการ และแต่ละการทดลองใช้เวลาอบเท่ากันคือ 15 นาที และหลังจากเอางานออกจากตู้อบความร้อนแล้วก็นำไปทำการทดลองต่อจากนั้นโดยทันที วิธีที่เขาใช้วัดความสามารถทนแรงดึงของกาวโดยใช้แรงดึงชิ้นงาน A และ B จนแยกออกจากกันได้ โดยที่การวัดค่าจะใช้วิธีค่อยๆเพิ่มแรงดึงทีละนิดจนทำให้ชิ้นงาน 2 ชิ้นนั้นแยกออกจากกันแล้วจดค่า แรงดึงสุดท้ายไว้ ผลการทดลองได้ค่าตามตารางนี้

70°C	80°C	90°C	100°C
2.3	2.5	3.0	3.3
2.6	2.9	3.1	3.5
2.1	2.4	2.8	3.0

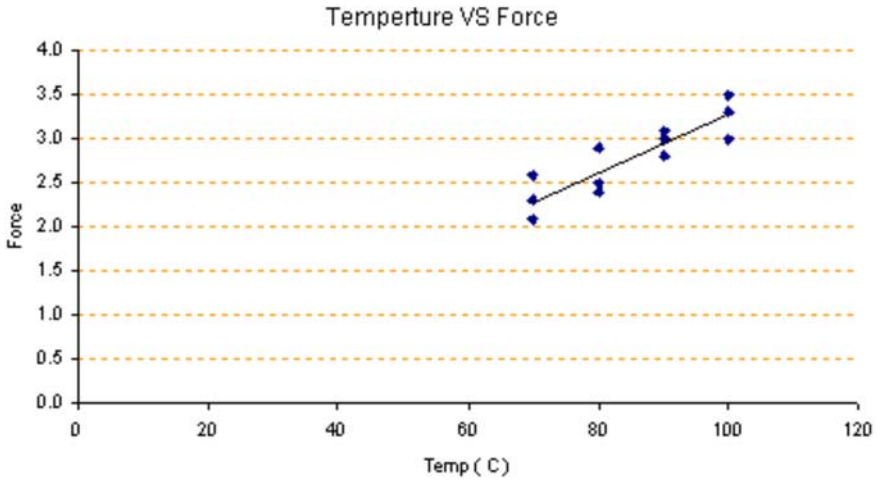
หน่วยคือ : หน่วยของแรง

เมื่อนำข้อมูลตามตารางมาทำ Scatter plot จะเป็นดังนี้



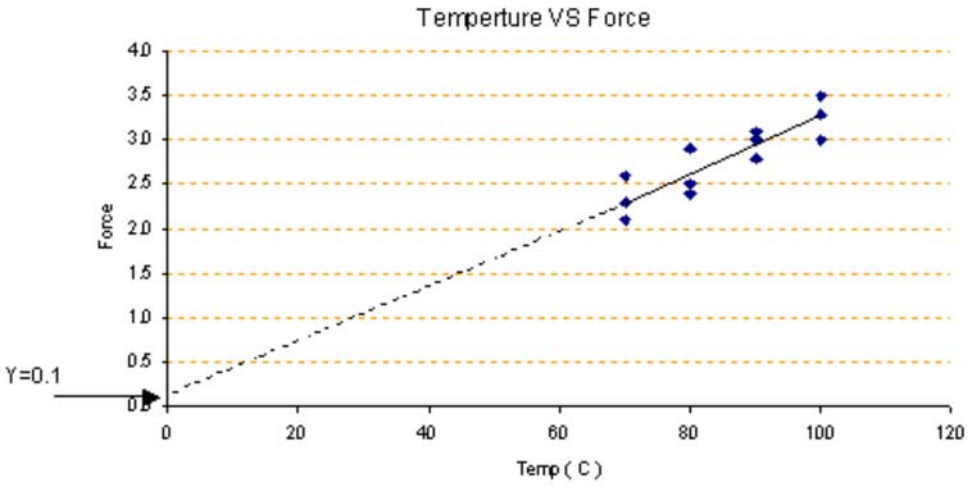
รูปที่ 1

จากกราฟ จะเห็นว่าค่าของตัวแปร Y (ตัวอย่างนี้คือ Force) ที่ค่า X ค่าเดิมนั้น จะมีค่าไม่เท่ากัน สมมติว่า แต่ละค่า X ผู้ทดลองเก็บค่า Y จำนวนมากๆ ผลที่ได้คือค่า Y ที่ X นั้นๆ ก็จะมีรูปแบบเป็น Normal distribution รอบค่ากลางค่าหนึ่ง(ค่าเฉลี่ย) และถ้าเราลากเส้นตรงเชื่อมกันระหว่างค่าเฉลี่ย ทุกจุดเข้าด้วยกันเราจะได้เส้นตรงเส้นหนึ่งที่เรียกว่า " Regression line" ตามรูปที่ 2



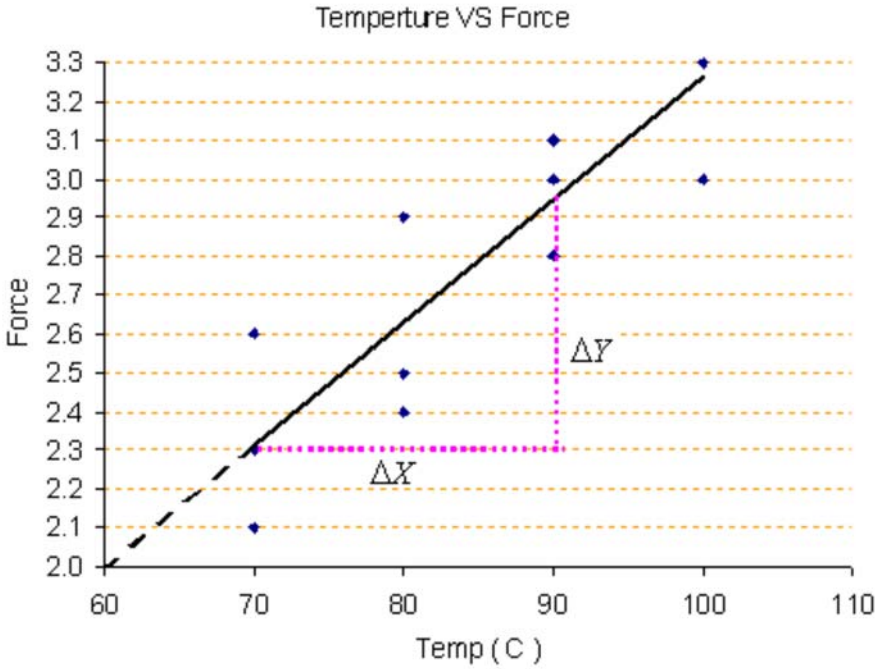
รูปที่ 2

ถ้าเราลากเส้นเชื่อมต่อมาจนตัดแกน Y จะตัดที่ค่า Y= 0.1 ค่านี้เราจะเรียกว่า Y-Intercept เขียนสัญลักษณ์แทนว่า b_0



รูปที่ 3

เมื่อเรานำแว่นขยายมาขยายจุดเหล่านี้เพื่อให้เห็นภาพใหญ่ขึ้นจะได้ดังรูปที่ 4 นี้



รูปที่ 4

เราสามารถหาค่าความชันของ Regression line ได้จาก

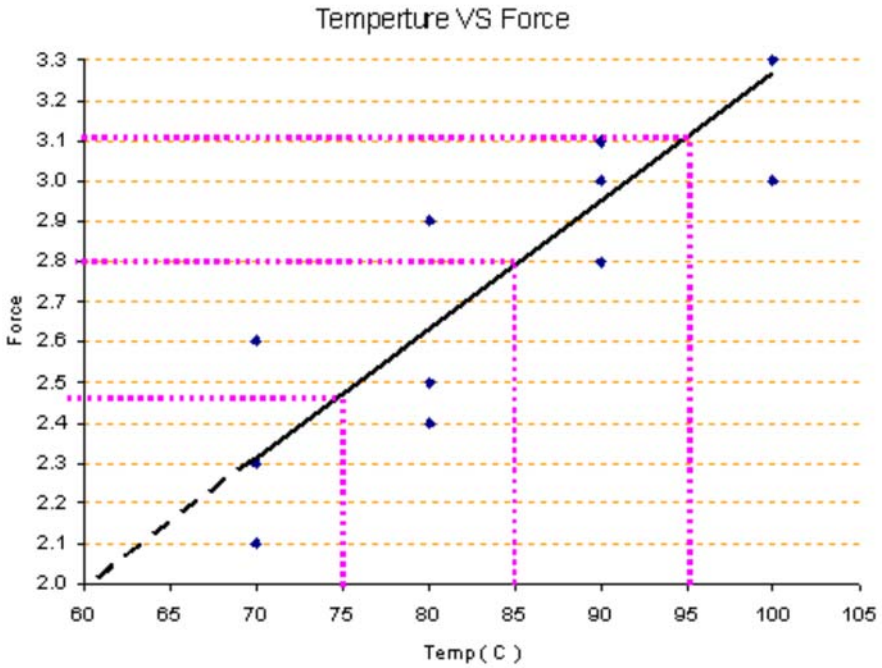
$$\frac{\Delta Y}{\Delta X} = \frac{2.95 - 2.3}{90 - 70} = 0.0325$$

ค่าความชันของ Regression line นี้เราจะแทนด้วยสัญลักษณ์ b_1 ดังนั้นเราจึงสามารถหาสมการของ Regression line ได้จาก

$$\hat{Y} = \beta_0 + \beta_1 X$$

$$\hat{Y} = 0.1 + 0.0325 X$$

หมายความว่า ณ ทุกจุดบนเส้นตรง (Regression line) นั้น ค่าในแนว แกน Y จะเท่ากับ $\hat{Y}$ เมื่อเราได้สมการแล้ว เราก็สามารถนำสมการนี้ไปใช้เพื่อ Predict ค่า Y เมื่อเรารู้ค่า X โดยเราจะลองใช้วิธีใช้เส้นตรง ดังต่อไปนี้



รูปที่ 5

นำค่าจากรูปที่ 5 ไปเขียนตาราง

Temperature	Force (Unit)
75°C	2.43
85°C	2.80
95°C	3.10

จะเห็นว่าค่าที่อยู่ในตารางนี้เป็นค่าที่ไม่ได้เกิดจากการทดลองจริงๆ แต่เป็นการเอาเส้นตรงที่ได้มาเป็นตัวช่วยในการคาดการณ์ (Prediction) เมื่อเป็นเช่นนี้ท่านผู้อ่านก็คงนึกถึงสมัยที่เราเรียนเรื่อง เส้นตรง และความชัน ในวิชาคณิตศาสตร์ชั้นมัธยมต้น ผู้เขียนจำได้ว่า เราใช้สมการ เส้นตรง และการหาค่าความชัน

ว่า

$$Y = mX + c$$

$$m = \frac{X_2 - X_1}{Y_2 - Y_1}$$

แต่อย่างที่ผู้เขียนได้เริ่มต้นเนื้อหาว่าค่าของตัวแปรในแนวแกน Y นั้นจะเป็นค่าที่มีการกระจายหรือมีความผิดพลาดโดยธรรมชาติ ดังนั้น Regression line คือเส้นที่ลากผ่านจุดๆหนึ่งของกลุ่มค่าแนวแกน Y โดยมีเงื่อนไขว่า ค่า Y ทั้งหมดจะห่างจากเส้นตรงนี้อย่างสมดุลกันมากที่สุด ไม่ใช่ค่า Y ทุกค่าอยู่บนเส้นนี้ นั่นแปลว่ายังต้องมีสิ่งหนึ่งที่ต้องคิดถึงคือค่าความห่าง ของค่า Y ใดๆ กับจุดบนเส้น Regression line ในแนวขนานกับเส้นแกน Y ค่าความห่างนี้เราเรียกว่า Error ใช้สัญลักษณ์แทนคือ e จากรูปที่ 5 ท่านจะเห็นว่า ค่า Y ใดๆ (เป็นจุด) แทบจะไม่มีค่าไหนอยู่บนเส้น Regression line เลย

ดังนั้นสมการนี้เมื่อนำไปใช้ก็จะได้ค่าความผิดพลาดมาด้วย ดังจะเห็นได้จากค่าที่ได้ในตาราง Predicting จะเห็นว่า ที่อุณหภูมิ 95 ยังได้ค่าเท่ากับที่ 90 องศา จุดหนึ่งด้วยซ้ำไป ดังนั้นสมการ เมื่อจะนำไป Predict ค่าจะต้องเป็น

$$\hat{Y} = \beta_0 + \beta_1 X + \varepsilon$$

ซึ่ง e มีค่าเฉลี่ยเท่ากับ 0 และมีค่า Variation เท่ากับ s^2 และเป็นสมการที่เกิดจากการวิเคราะห์ Yi และ Xi เพียงจุดใดๆ เท่านั้น

แต่การหาค่า *b*0 และ *b*1 ตามวิธีที่ผ่านมานั้นเป็นการใช้กราฟ อาจจะทำให้เราได้ค่าที่ผิดพลาดไปบ้างอันเนื่องจากการเทียบค่าจาก Regression line มายังแกน X และ Y อาจมีความคลาดเคลื่อนไป ในทางปฏิบัติเราจึงไม่นิยมนำมาใช้ในการประมาณค่า โดยจะนิยมใช้วิธีการที่เรียกว่า " Method of least square" มากกว่า วิธีที่ว่านี้เป็นการรวมวิเคราะห์ จุด Xi และ Yi ทุกๆจุดเพื่อหาค่า ซึ่งมีวิธีดังนี้

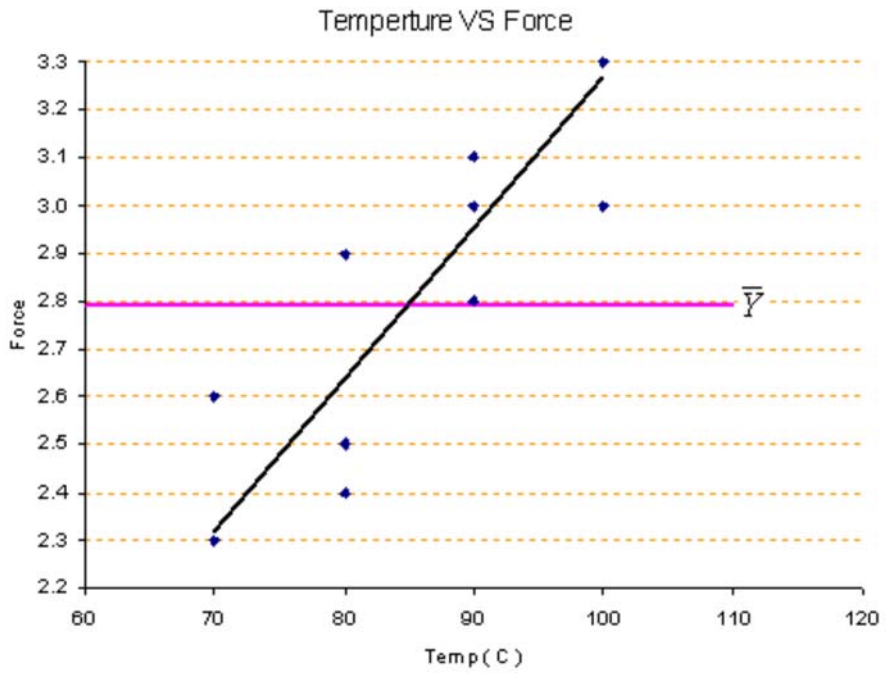
1. นำค่าจากผลการทดลองมาสร้างตารางใหม่ดังต่อไปนี้

n	x	y	x^2	xy	y^2
1	70	2.3	4900	161	5.29
2	70	2.6	4900	182	6.76
3	70	2.1	4900	147	4.41
4	80	2.5	6400	200	6.25
5	80	2.9	6400	232	8.41
6	80	2.4	6400	192	5.76
7	90	3.0	8100	270	9.00

8	90	3.1	8100	279	9.61
9	90	2.8	8100	252	7.84
10	100	3.3	10000	330	10.89
11	100	3.5	10000	350	12.25
12	100	3.0	10000	300	9.00
Sum	1020	33.5	88200	2895	95.47

ที่เราต้องทำเพิ่มขึ้นคือ หาค่า x^2 , xy และ y^2 ในตาราง

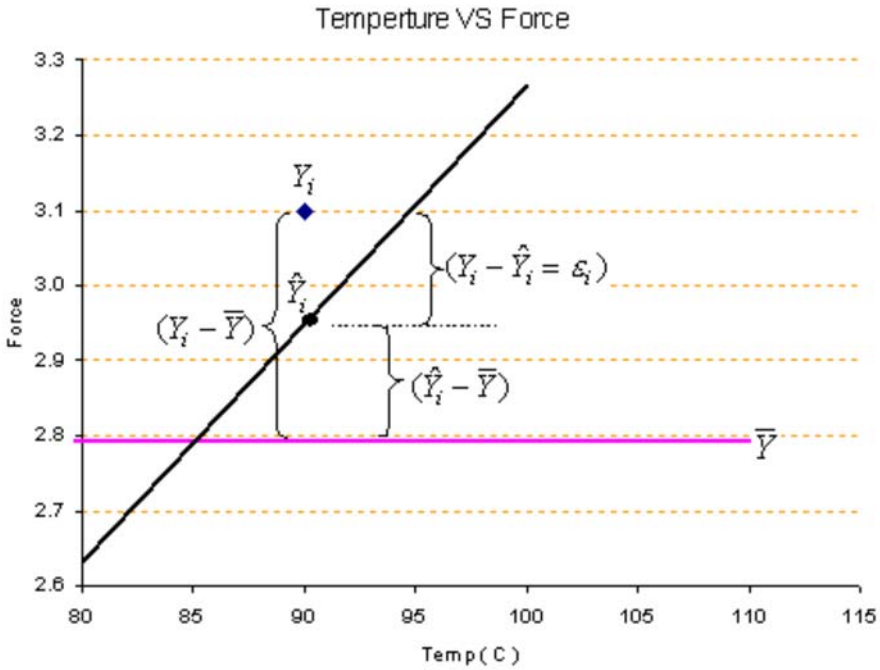
2. หาค่าเฉลี่ยของค่าที่วัดได้ทั้งหมดทุกค่า จากตัวอย่างนี้ก็คือการนำค่า Y ทั้งหมดในตารางมาหาค่าเฉลี่ย ซึ่งเราจะได้ค่า $\bar{Y}$



รูปที่ 6

นั่นคือ ถ้าเราไม่ใช้ X ในการคาดการค่า Y แล้ว ค่าโดยเฉลี่ย $\overline{Y}$ จะเป็นค่าที่ใช้คาดการณ์ Y ได้ดีที่สุด

3. วิเคราะห์ แต่ละจุด โดยเทียบกับค่า $\overline{Y}$ และ $\hat{Y}$ เพื่อให้เห็นภาพและเข้าใจง่ายขึ้น ผู้เขียนจะวิเคราะห์ค่า Yi เพียง 1 จุดให้เห็นเป็นตัวอย่างดังนี้



รูปที่ 7

จากข้อ 2 และ 3 นั้นชี้ให้เห็นว่า ในแนวแกน Y เราจะมองค่า Yi ของทุกตัวเป็นค่าเฉลี่ย ($\overline{Y}$) เพียง 1 ค่า และในแนวแกน X เราจะมองค่า Yi ของทุกตัวเพียง 1 ค่าอยู่บนเส้น Regression line ($\hat{Y}$) เท่านั้น การมองค่า Yi ใน 2 แกนเช่นนี้ ภาษาอังกฤษ เขาเรียกว่า " Regress " ถ้าความหมายเป็นไทยก็น่าจะตรงกับความหมายว่า " การมองค่า Yi ทุกค่าย้อนกับไปสู่จุดรวมใน 2 แนว " และนี่เป็นที่มาของคำว่า Regression ซึ่งการที่เรามองวิธีการคิดแบบนี้เข้าใจแล้ว การที่เราเรียกว่า "ถดถอย" ไม่น่าจะถูกต้อง ในความคิดส่วนตัวของผู้เขียน แต่บังเอิญถ้าเปิด Dictionary ก็คงจะพบคำแปลเป็นไทย อย่างนั้น ก็เลยเรียกกันอย่างนั้น

จากตาราง จะได้

$$\bar{x} = 85, \bar{y} = 2.79$$

$$\beta_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - n(\bar{x})^2}$$

$$\beta_1 = \frac{2895 - \frac{(1020)(33.5)}{12}}{88200 - (12)(85)^2}$$

$$\beta_1 = \frac{47.5}{1500} = 0.032$$

$$\beta_0 = \bar{y} - \beta_1(\bar{x}) = 2.79 - (0.032)(85) = 0.07$$

$$\hat{y} = 0.07 + 0.32x$$

*** หมายเหตุ ค่าที่ได้อาจแตกต่างจากวิธีคำนวณโดยวิธีอื่นๆเล็กน้อย เนื่องจากการปัดเศษ หรือทศนิยม*

สมการที่ได้นี้คือ Regression หรือคือค่าในแนวแกน Y ทุกจุด บนเส้น Regression line จะมี่ค่าตามสมการที่ได้นี้ เมื่อเรารู้ค่า x เราก็สามารถรู้ค่า Response หรือค่าในแนวแกน Y ได้โดยการลากเส้นตั้งฉากกับ แกน Y จากจุดบนแกน X มาตัดกับ Regression line เราก็จะทราบได้ทันที

[HOME] [CONTENTS]

ตัวอย่างการคำนวณ Simple Linear Regression Analysis

ตัวอย่าง คณะนักวิจัยที่ทำการศึกษาเกี่ยวกับชีวิตของหมีป่าต้องการหาวิธีการจะประมาณค่าน้ำหนักของหมีที่อาศัยอยู่ตามธรรมชาติในป่า โดยมีจุดประสงค์ที่สำคัญคือในอนาคตข้างหน้านักวิจัยกลุ่มนี้ไม่ต้องการใช้เครื่องชั่งน้ำหนักในการทำการศึกษาเกี่ยวกับชีวิตหมีอีก ทั้งนี้เพราะมีความลำบากในการขนย้ายเหลือเกิน จึงทำการออกสำรวจกลุ่มตัวอย่างหมีป่า 10 ตัว หลากหลายขนาด ทั้งเพศผู้และเพศเมีย แล้วทำการชั่งน้ำหนักและวัดความยาวรอบบริเวณราวนมของหมีป่า ทั้ง 10 ตัว ได้ข้อมูลดังตารางต่อไปนี้

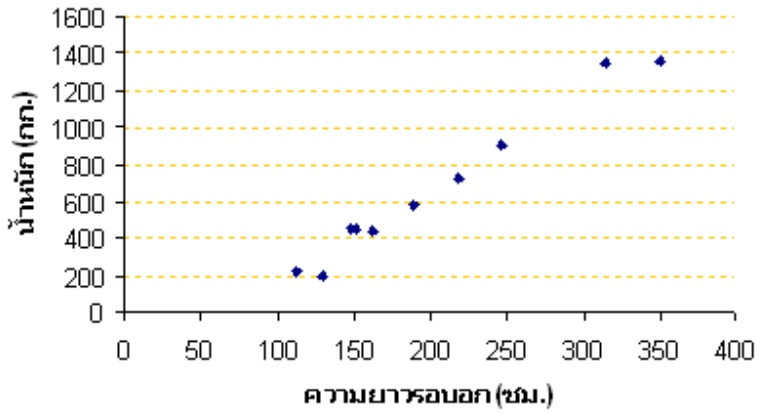
หมีตัวที่	1	2	3	4	5	6	7	8	9	10
รอบอก (ซม.)	112	130	148	151	162	189	218	247	315	350
น้ำหนัก (กก.)	225	200	459	445	439	577	722	903	1350	1360

ตารางที่ 1.

เมื่อมีตัวแปรต้นและตัวแปรตามเพียงอย่างละ 1 ตัวแปร สมการหรือ Model ที่ได้จะเป็นดังนี้

$$Y = \beta_0 + \beta_1 X$$

ขั้นตอนที่ 1 เตรียมข้อมูล X, Y จากตาราง ค่าที่เป็นตัวแปรต้นหรือ X คือความยาวเส้นรอบอก และค่าที่เป็นตัวแปรตามหรือ Y ก็คือค่าน้ำหนักนั่นเอง ซึ่งต้องเลือกให้ตรงด้วย หากท่านผู้อ่านสงสัยว่าเหตุใดผู้เขียนถึงเลือกแบบนี้ ที่จริงไม่ใช่เลือกครับแต่ความเป็นจริงต่างหาก วิธีสังเกตคือค่าที่เราจะหาหรือจะเป็นคำตอบในอนาคตคือตัวแปรตาม (Y) ส่วนค่าที่เราจะต้องใส่ในสมการ (Model) เพื่อให้เห็นคำตอบคือตัวแปรต้น ในโจทย์นี้ค่าที่จะเป็นคำตอบจากการใช้ Model คือค่าน้ำหนักนั่นเอง



เมื่อเอาข้อมูล X,Y มาทำ Scatter plot พบว่าการเรียงตัวของจุดมีลักษณะเป็นเส้นตรง พอจะมองเห็นว่ามีความสัมพันธ์เป็นแบบเชิงเส้น และแนวของจุดชันขึ้นทำมุมกับแกน X มากพอประมาณ แสดงว่าค่าสัมประสิทธิ์ของตัวแปรต้นมากกว่า 0 หรือ Slope > 0 เข้าข่ายที่จะพิสูจน์ด้วย Regression analysis ได้

ขั้นตอนที่ 2 หาค่าผลรวม ค่าเฉลี่ย ผลคูณ ผลรวมของการคูณกันของตัวแปรต้นและตัวแปรตาม และค่ายกกำลังสองของตัวแปรต้น ท่านผู้อ่านต้องใช้โปรแกรม Excel ช่วยในการคำนวณ โดยคัดลอกข้อมูลจากตารางไปวางในตาราง Excel จากการคำนวณจะได้ค่าต่างๆดังตารางต่อไปนี้

	Sum	Average

X	112	130	148	151	162	189	218	247	315	350	2022	202.2
Y	225	200	459	445	439	577	722	903	1350	1360	6680	668
XY	25200	26000	67932	67195	71118	109053	157396	223041	425250	476000	1648185	
X^2	12544	16900	21904	22801	26244	35721	47524	61009	99225	122500	466372	

ตารางที่ 2.

ขั้นตอนที่ 3 คำนวณหาค่าคงที่และค่าสัมประสิทธิ์ของตัวแปรต้นในสมการ โดยนำค่าที่คำนวณได้จากตารางที่ 2 มาใส่ในสมการ ซึ่งสมการคำนวณดังกล่าวนี้ได้มีการคิดไว้แล้ว ผู้เขียนขอไม่พูดถึงที่มาของสมการในขั้นตอนนี้

$$\beta_1 = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - n(\overline{X})^2}$$

$$= \frac{1648185 - \frac{(2022)(6680)}{10}}{466372 - 10(202.2)^2}$$

$$= 5.1716$$

$$\beta_0 = \overline{Y} - \beta_1(\overline{X})$$

$$= 668 - 5.1716(202.2)$$

$$= -377.7$$

ดังนั้น Regression model ที่ได้ คือ

$$Y = -377.7 + 5.17(X)$$

หรือเขียนเป็นสมการความสัมพันธ์ตามข้อมูลจากตารางแรก คือ

น้ำหนักหมี(โดยประมาณ) = -377.7 + 5.17 (ขนาดเส้นรอบวงบริเวณอก ซม.) กก.

ขั้นตอนที่ 4 ทดสอบสมมติฐานว่าค่าคงที่ (b_0) และค่าสัมประสิทธิ์ (b_1) ใน Model ที่คำนวณได้นั้นมีความจำเป็นต้องคงไว้ใน Model หรือไม่ ท่านผู้อ่านต้องนึกภาพตามนะครับว่าค่า b_0 และ b_1 ที่เรากำลังกล่าวถึงนี้เป็นตัวอย่างที่มาจากประชากรจำนวนหนึ่ง ซึ่งจะมีการกระจายรอบๆค่าๆหนึ่ง ซึ่งจริงๆแล้วค่าที่เราเห็นก็เป็นเพียงตัวอย่างหนึ่งที่ถูกดึงออกมาจากกลุ่ม ที่เราสนใจคือว่าค่าดังกล่าวนี้ เป็น 0 หรือไม่ ถึงเราจะเห็นค่าไม่เท่ากับ 0 แต่จริงๆค่าการกระจายกลุ่มนั้นอาจจะอยู่รอบๆ 0 ก็ได้ ถ้าเป็นเช่นนั้นถือว่าค่าดังกล่าวไม่มีนัยสำคัญของความต่างกับ 0 เราก็สามารถตัดค่า b_0 หรือตัดพจน์ที่ b_1 คูณอยู่ออกจาก Model ได้เลย โดยจะไม่ทำให้ค่า Y ที่ได้มีความแตกต่างกับเมื่อคงค่าดังกล่าวไว้ใน Model แต่อย่างใด และการกระจายของตัวของค่า b_0 และ b_1 นี้ก็เป็น Normal distribution ด้วย เหตุผลนี้เองที่เราต้องใช้ T-Test เพื่อทดสอบสมมติฐาน ว่ามีค่าเท่ากับ 0 หรือไม่

โดยสมมติฐานในการทดสอบ b_0 คือ

$$H_0 : \beta_0 = 0$$
$$H_a : \beta_0 \neq 0$$

การพิสูจน์สมมติฐานนั้นเราจะใช้ t-statistic โดยมีสมการดังนี้

$$\beta_0(i) = Y(i) - \beta_1 X(i)$$
$$(\sigma_{\beta_0})^2 = \frac{\sum(\beta_0(i) - \beta_0)^2}{n}$$
$$t_{\beta_0} = \frac{\beta_0}{\sigma_{\beta_0}}$$

จากสมการข้างบน เราจะเริ่มจากการหาค่า $b_{0(i)}$ และ s_{b0} เพื่อจะนำไปใช้หาค่า t_{b0} ในขั้นตอนสุดท้าย ผู้เขียนขอใช้ตาราง Excel ช่วยในการคำนวณ จาก Model ค่า b_0 = -377.7 และ b_1 = 5.17

X (i)	Y (i)	$b_{0\,(i)}$	$(b_{0\,(i)} - b_0)$	$(b_{0\,(i)} - b_0)^2$
112	225	-354.22	23.48	551.23
130	200	-472.31	-94.61	8951.15
148	459	-306.40	71.30	5083.79
151	445	-335.91	41.79	1746.06
162	439	-398.80	-21.10	445.28
189	577	-400.43	-22.73	516.88
218	722	-405.41	-27.71	767.91
247	903	-374.38	3.31	10.97
315	1350	-279.05	98.64	9730.56
350	1360	-450.06	-72.36	5236.31
			Sum	**33040.14**
			$(s_{b0})^2$	**3304.01**

ตารางที่ 3.

จากตารางที่ 3 จะได้

$$t_{\beta_0} = \frac{\beta_0}{\sigma_{\beta_0}} = \frac{-377.7}{\sqrt{3304.01}} = \frac{-377.7}{57.48} = -6.57$$

เราจะปฏิเสธสมมติฐาน $H_0 : b_0 = 0$ ถ้า

$$|t_{\beta_0}| > t_{\alpha/2,n-2}$$

(α = 0.05 , df = n-2)

จากตาราง T เราพบว่า $t_{0.025,8}$ = 2.306 ซึ่งน้อยกว่า $|t_{b0}|$ ที่คำนวณได้ ดังนั้นสมมติฐาน $H_0 : b_0 = 0$ จึงไม่เป็นจริง นั่นคือค่า b_0 มีค่ามากกว่า 0 อย่างมีนัยสำคัญ จึงต้องคงค่าไว้ใน Model

สมมติฐานในการทดสอบ b_1 คือ

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

การพิสูจน์สมมติฐานนั้นเราจะใช้ t-statistic โดยมีสมการดังนี้

$$(\sigma_{\beta_1})^2 = \frac{\frac{\sum (Y(i) - \hat{Y})^2}{n-2}}{\sum (X(i) - \bar{X})^2}$$

$$t_{\beta_1} = \frac{\beta_1}{\sigma_{\beta_1}}$$

เมื่อ $\hat{Y}$ คือค่าที่ได้จากการนำ Model ที่ได้มานั้นมาใส่ค่า X แล้วหาค่า Y

X	Y	$X(i) - \bar{X}$	$(X(i) - \bar{X})^2$	$\hat{Y}$	$Y(i) - \hat{Y}$	$(Y(i) - \hat{Y})^2$
112	225	-90.2	8136.04	201.34	23.66	559.7956
130	200	-72.2	5212.84	294.4	-94.4	8911.36
148	459	-54.2	2937.64	387.46	71.54	5117.972
151	445	-51.2	2621.44	402.97	42.03	1766.521
162	439	-40.2	1616.04	459.84	-20.84	434.3056
189	577	-13.2	174.24	599.43	-22.43	503.1049
218	722	15.8	249.64	749.36	-27.36	748.5696
247	903	44.8	2007.04	899.29	3.71	13.7641
315	1350	112.8	12723.84	1250.85	99.15	9830.723
350	1360	147.8	21844.84	1431.8	-71.8	5155.24
Average	**202.2**					
Sum			**57523.6**			**33041.35**

ตารางที่ 4.

จากตารางที่ 4 จะได้

$$(\sigma_{\beta_1})^2 = \frac{\frac{33041.35}{8}}{57523.6} = \frac{4130.17}{57523.6} = 0.0718$$

$$\therefore\ t_{\beta_1} = \frac{5.17}{\sqrt{0.0718}} = 19.30$$

เราจะปฏิเสธสมมติฐาน $H_0 : b_1 = 0$ ถ้า

$$|t_{\beta_1}| > t_{\alpha/2, n-2}$$

($a = 0.05$, $df = n-2$)

จากตาราง T เราพบว่า $t_{0.025,8} = 2.306$ ซึ่งน้อยกว่า $|t_{b_1}|$ ที่คำนวณได้ ดังนั้นสมมติฐาน $H_0 : b_1 = 0$ จึงไม่เป็นจริง นั่นคือค่า b_1 มีค่ามากกว่า 0 อย่างมีนัยสำคัญ จึงต้องคงพจน์ที่มี b_1 เป็นสัมประสิทธิ์คูณอยู่ไว้ใน Model ไม่สามารถตัดทิ้งได้

นั่นคือ Regression model ยังคงเป็น

$$Y = -377.7 + 5.17(X)$$

ขั้นตอนที่ 5 การพิสูจน์ว่า Regression model ที่ได้มานั้นเหมาะที่จะนำไปใช้คาดการณ์ (Predict) ค่า Y ในอนคตมากน้อยเพียงใด ซึ่งจะใช้วิธีพิสูจน์ค่าความคลาดเคลื่อน (Error) ระหว่างค่า Y ที่เก็บข้อมูลมาได้กับค่าที่ได้จากการใส่ค่า X ใน Model ที่ได้มา ($\hat{Y}$) ซึ่งเรียกว่า Residual นั่นเอง ตัวสถิติที่จะใช้ทดสอบความคลาดเคลื่อนนี้ เราเรียกว่า F-Statistic และสมมติฐานคือ

H_0 : Error จากการใช้ Model นี้ Predict ค่า Y เป็น Error ที่ไม่สามารถอธิบายได้เป็นส่วนใหญ่

H_a: Error จากการใช้ Model นี้ Predict ค่า Y เป็น Error ที่สามารถอธิบายได้เป็นส่วนใหญ่

สมการทางคณิตศาสตร์ที่ใช้ในการคำนวณ มีดังนี้

$$SS_{Error} = \sum \left(Y(i) - \hat{Y}\right)^2$$

$$SS_{Total} = \sum \left(Y(i) - \bar{Y}\right)^2$$

$$SS_{Regression} = SS_{Total} - SS_{Error}$$

$$MS_{Regression} = \frac{SS_{Regression}}{df}$$

$$MS_{Error} = \frac{SS_{Error}}{df}$$

$$F = \frac{MS_{Regression}}{MS_{Error}}$$

โดยที่

SS : Sum Square หมายถึงค่าแต่ละค่ายกกำลังสองแล้วนำมาหาผลรวม

MS : Mean Square หมายถึงค่าการเอาค่า SS มาหาค่าเฉลี่ยอีกโดยหารด้วย Degree of freedom.

X	Y	$Y(i) - \bar{Y}$	$(Y(i) - \bar{Y})^2$	$\hat{Y}$	$Y(i) - \hat{Y}$	$(Y(i) - \hat{Y})^2$
112	225	-443	196249	201.34	23.66	559.7956
130	200	-468	219024	294.4	-94.4	8911.36
148	459	-209	43681	387.46	71.54	5117.972
151	445	-223	49729	402.97	42.03	1766.521
162	439	-229	52441	459.84	-20.84	434.3056
189	577	-91	8281	599.43	-22.43	503.1049
218	722	54	2916	749.36	-27.36	748.5696
247	903	235	55225	899.29	3.71	13.7641
315	1350	682	465124	1250.85	99.15	9830.723
350	1360	692	478864	1431.8	-71.8	5155.24
Average		**668**				
Sum			**1571534**			**33041.35**

ตารางที่ 5.

จากตารางที่ 5

SS_{Error} = 33041.35

SS_{Total} = 1571534

ดังนั้น

$SS_{Regression}$ = 1571534 - 33041.35 = 1538492.645

หาค่า Degree of freedom

SS_{Total} : df = n -1 =10 -1 = 9

SS_{Error} : df = n -1-1 =10 -1 -1 = 8 >> มาจาก (df_{Total} - 1)

$SS_{Regression}$: df = n -1 - 8 =10 -1-8 = 1 >> มาจาก (df_{Total} - df_{Error})

ดังนั้น

$$MS_{Regression} = \frac{SS_{Regression}}{df} = \frac{1538492.645}{1} = 1538492.645$$

$$MS_{Error} = \frac{SS_{Error}}{df} = \frac{33041.35}{8} = 4130.17$$

$$F = \frac{MS_{Regression}}{MS_{Error}} = \frac{1538492.645}{4130.17} = 372.5011$$

เราจะปฏิเสธสมมติฐาน H_0 ถ้า $F > F_{Critical}$ ($\alpha = 0.05$, df >> $n_1=1$, $n_2=8$)

จากตาราง F เราพบว่า $F_{0.05,1,8} = 5.32$ ซึ่งน้อยกว่า F ที่คำนวณได้ ดังนั้นสมมติฐาน H_0 จึงไม่เป็นจริง นั่นคือ Error จากการใช้ Model นี้ Predict ค่า Y เป็น Error ที่สามารถอธิบายได้เป็นส่วนใหญ่ หมายความว่าความแตกต่างของค่า Y ที่เห็นส่วนใหญ่เกิดจากการใส่ค่า X ที่ต่างกัน ซึ่งเป็นสิ่งที่อธิบายได้ ส่วนเหตุอื่นๆที่มีผลทำให้เกิดความแตกต่างของค่า Y ในการใช้ Model นี้ มีน้อยมาก

เราใช้ F -Statistics เพื่อรับรองว่า หากใช้ Regression model นี้ไปใช้ Predict ค่า Y แล้วจะให้ความคลาดเคลื่อนน้อย หรือพูดอีกอย่างคือมีความแม่นยำสูงนั่นเอง

ขั้นตอนที่ 6 การหา Coefficient of Determination เป็นการคำนวณหาตัวชี้วัดว่า Model นี้สมควรจะได้รับการยอมรับมากน้อยเพียงใด ถึงแม้จะรับรองด้วย F-Statistics แล้วก็ตาม หลักการคือหาค่า Error จากการเปลี่ยนแปลงค่า X ซึ่งเป็นการเปลี่ยนแปลงที่เราจงใจ กับค่า Error รวมทั้งหมด ถ้าค่าที่ได้ใกล้เคียงกัน ก็ถือว่า ยอมรับได้ ถ้าน้อย ก็แสดงว่าค่า Error อื่นๆที่ไม่รู้ที่ไปที่มา มีปนอยู่มากถึงระดับหนึ่งอาจจะไม่สามารถยอมรับ Model นี้ได้เลย เราเรียกตัวชีวัดนี้ว่า R^2 (อ่านว่า R - Square)

$$R^2 = \frac{(SS_{Regression})100}{SS_{Total}}\%$$

$$R^2 = \frac{1538492.645}{1571534} = 97.9\%$$

อาจจะเป็นไปได้ว่าเพราะความบังเอิญค่า R^2 ที่คำนวณได้จึงสูง เราจะต้องทดสอบดูว่า ที่ค่า R^2 สูงนั้นไม่ได้เป็นเรื่องบังเอิญ หลักการคือให้ทำการลด n ลง 1 ตัวแล้วหาค่า R^2 อีกครั้ง หากยังสูงอยู่ก็ถือว่าไม่ได้เป็นเรื่องบังเอิญ แต่ถ้า R^2 ใหม่นี้มีค่าต่ำกว่าค่าเดิมมาก แสดงว่าค่า R^2 มีความไว (Sensitivity) ต่อการเปลี่ยนแปลง n มาก ควรจะต้องแก้ไข โดยอาจถึงขั้นต้องไปเก็บข้อมูลเพิ่ม เก็บข้อมูลใหม่ เลยทีเดียว เราเรียกว่า R^2 -Adjusted

$$R^2 - Adjusted = \left\{1 - \left(\frac{n-1}{n-p}\right)(1-R^2)\right\}100\%$$

$$= \left\{1 - \left(\frac{10-1}{10-2}\right)(1-0.979)\right\}100\%$$

$$= 97.63$$

โดยที่ p คือจำนวนค่าคงที่และค่าสัมประสิทธิ์ของตัวแปรต้นใน Regression model ($b_0, b_1, b_2, \ldots b_n$) ซึ่งในตัวอย่างนี้คือ 2 คือ b_0, b_1 นั่นเอง

หากนำค่าที่ได้จากการคำนวณมาเขียนสรุปเป็นตารางจะได้ดังต่อไปนี้

b_0 =	-377.7	t = -6.57	F = 372.5
b_1 =	5.17	t = -19.3	
R^2 =	0.979		
R^2-Adjusted =	0.9763		

ตารางที่ 6.

http://stattrek.com/AP-Statistics-4/Test-Slope.aspx?Tutorial=AP

ในกรณีที่เราใช้โปรแกรม Microsoft Excel ช่วยในการวิเคราะห์ จะได้ตารางออกมาดังต่อไปนี้

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.989
R Square	0.979
Adjusted R Square	0.976
Standard Error	64.265
Observations	10

$$SE = s_{b_1} = \text{sqrt}\left[\Sigma(y_i - \hat{y}_i)^2 / (n-2)\right] / \text{sqrt}\left[\Sigma(x_i - \bar{x})^2\right]$$

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1538493.86	1538493.86	372.52	0.00000
Residual	8	33040.14	4130.02		
Total	9	1571534			

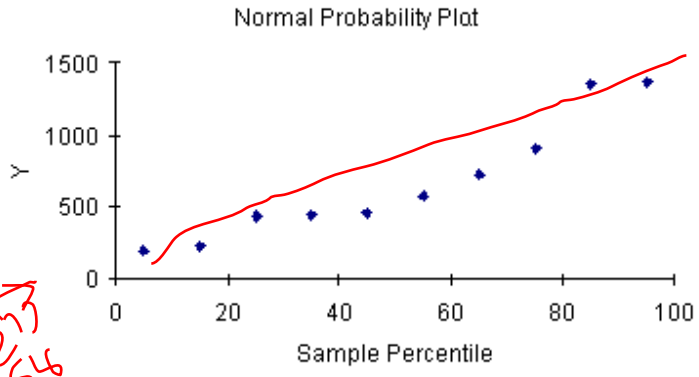
	Coefficients	Standard Error	t Stat	P-value
Intercept	-377.70	57.87	-6.53	0.00
X	5.17	0.27	19.30	0.00

ตารางที่ 7.

จะเห็นว่าถ้าเราใช้โปรแกรมช่วยวิเคราะห์ เราจะเห็นตารางค่าคงที่ ค่าสัมประสิทธิ์ของตัวแปรต้น และค่า T-Statistics (ตารางสีเทา) ซึ่งได้ค่า P-Value เป็น 0.00 แสดงว่าเราต้องปฏิเสธสมมติฐาน H_0: ในขั้นตอนที่ 4 ทั้งกรณีสมมติฐานของค่า b_0 และ b_1 ในส่วนตาราง Anova (ตารางสีชมพู) ใช้เพื่อหาค่า F Statistics ได้ค่า P-Value เป็น 0.00 นั่นคือปฏิเสธสมมติฐาน H_0 : ตามขั้นตอนที่ 5 เช่นกัน ในส่วนตาราง Regression statistics (ส่วนสีฟ้าอ่อน) ที่สรุปค่า Coefficient of Determination ที่ได้ตามขั้นตอนที่ 6 สุดท้ายเราสรุปว่าค่าที่คำนวณตามขั้นตอนที่ 1 ถึง 6 ได้ตรงกับค่าที่ได้จากการใช้โปรแกรม Micrsoft Excel

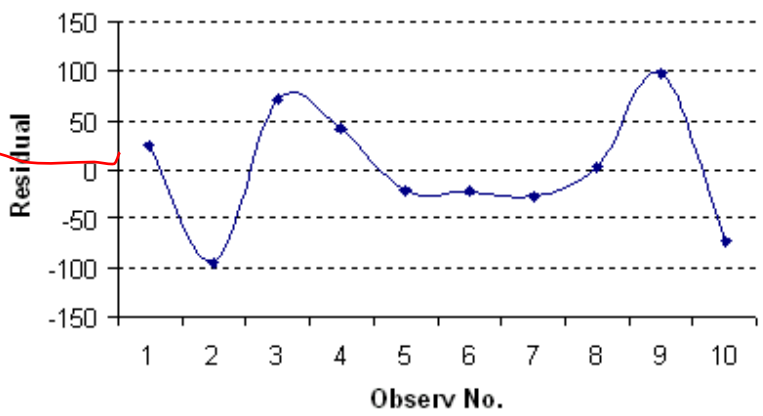
ขั้นตอนที่ 7 การพิสูจน์คุณสมบัติ 3 ประการ เนื่องจากในการคำนวณด้วยมือตามขั้นตอนที่ 1-6 มีความยุ่งยากในการเตรียมข้อมูลในการแสดงกราฟ ผู้เขียนขอใช้กราฟที่ได้จากการวิเคราะห์ด้วยโปรแกรม Mocrosoft Excel

- Normality



จากกราฟ การเรียงตัวของจุดค่า Y เทียบกับ Percentile เป็นแนว แม้จะไม่เป็นเส้นตรง โอกาสที่จะไม่เป็น Normal ค่อนข้างสูงทีเดียว แต่หากจะยอมรับว่าเข้าใกล้ Normal distribution ก็ไม่น่าเกลียดเกินไป

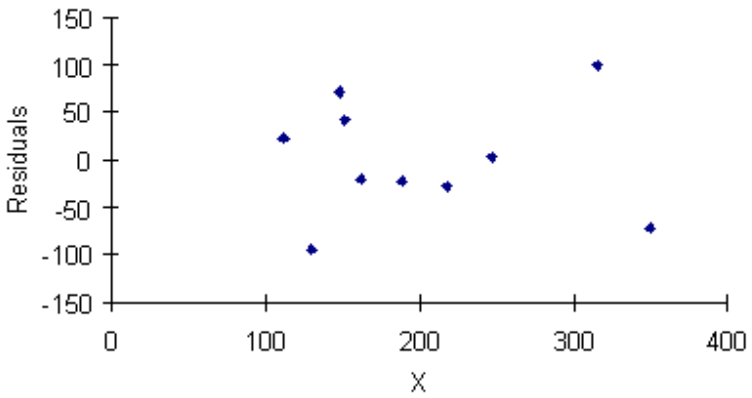
- Independence



การทดสอบ Independence หรือความเป็นอิสระต่อกันของค่า X แต่ละค่าทำได้โดยการพล้อตค่า Residual เทียบกับหมายเลขของ X ที่เรากำหนดในตาราง จะพบว่า แนวของจุดถือได้ว่า ไม่มีทิศทางใดแน่นอน ไม่ได้อยู่ทางด้านลบหรือบวกอย่างเดียว ไม่ได้ขึ้น

หรือลงอย่างเดียว ลักษณะเช่นนี้เราถือว่าความเป็นอิสระของ X แต่ละตัวอยู่ในเกณฑ์ที่ยอมรับได้

- Homoscedasticity



การทดสอบ Homoscedasticity มีวัตถุประสงค์คือ พิสูจน์ว่าค่าความคลาดเคลื่อนทุกๆย่านของค่า X ไม่ได้แตกต่างกันมากจนเกินไป โดยการพล้อต Residual กับค่า X (Fit) หากมีลักษณะอยู่ด้านบวก หรือลบตลอด เป็น 0 ตลอด กว้างออกตลอด เมื่อค่า X สูงขึ้น เราจะถือว่าไม่ผ่านเงื่อนไขนี้ จากกราฟเราพอจะอนุมานได้ว่า Residual ตลอดย่านค่า X ไม่ได้แตกต่างกันจนเกินเหตุ

ดังนั้น เงื่อนไขที่สำคัญทั้ง 3 ก็ถือว่ายอมรับได้ เราจึงยอมรับว่า Regression model ที่ได้มานั้นสามารถเอาไปใช้ในการคาดการณ์ค่าน้ำหนักหมีป่าในอนาคต โดยจะได้คำตอบโดยประมาณที่ใกล้เคียงกับความเป็นจริงพอสมควร ซึ่งเพียงพอที่นักวิจัยจะใช้ค่าน้ำหนักที่คำนวณได้ไปใช้ในการพิสูจน์หรือวิเคราะห์อะไรก็ตามที่ต้องใช้ค่าน้ำหนักหมีมาเกี่ยวข้อง แล้วยังให้ข้อสรุปที่ถูกต้องยอมรับได้อยู่ เช่น ถ้าต่อไปนักวิจัยกลุ่มนี้ไปสำรวจหมีตัวที่ 11 พบว่ามีเส้นรอบอก 275 ซม. เมื่อใส่ค่าใน Regression model ก็จะได้ค่าน้ำหนักหมีประมาณ 1044 กก. อย่าลืมว่าคำตอบที่ได้ต้องเป็นค่าประมาณการณ์เท่านั้น

ถึงแม้ขั้นตอนการวิเคราะห์ Regression จะดูยาวและต้องทำหลายอย่าง แต่เนื่องจากในปัจจุบันนี้ เรามีโอกาสใช้โปรแกรมคอมพิวเตอร์ช่วยในการวิเคราะห์ โดยที่เราไม่ต้องทำเอง ซึ่งก็จะได้ค่าเป็นตารางรายงานและกราฟออกมาให้เห็นเลย แต่ถ้าหากเราไม่เข้าใจว่าการวิเคราะห์หรือการคำนวณมีที่ไปที่มาอย่างไร ค่าที่เห็นในตารางที่คอมพิวเตอร์ให้มานั้น แต่ละค่ามาอย่างไร กราฟแต่กราฟหมายถึงอะไร และจะตีความอย่างไรดี ผู้เขียนก็ขอบอกว่าคอมพิวเตอร์ก็ช่วยไม่ได้ แต่เมื่อท่านผู้อ่านเข้าใจขั้นตอนวิเคราะห์แต่ละค่าในตารางมีสมการในการคิดอย่างไรอยู่ กราฟแต่ละกราฟใช้ค่าอะไรพล้อต ใช้ดูอะไรแล้ว ก็จะสามารถอ่านและสรุปผลจากคอมพิวเตอร์ได้ถูกต้อง แม้แต่หากใส่ข้อมูลผิดก็ยังสามารถมองเห็นข้อผิดพลาดได้ คอมพิวเตอร์ก็มีความหมายเพียงเครื่องช่วยประมวลผลให้เราเท่านั้น

[HOME] [CONTENTS]

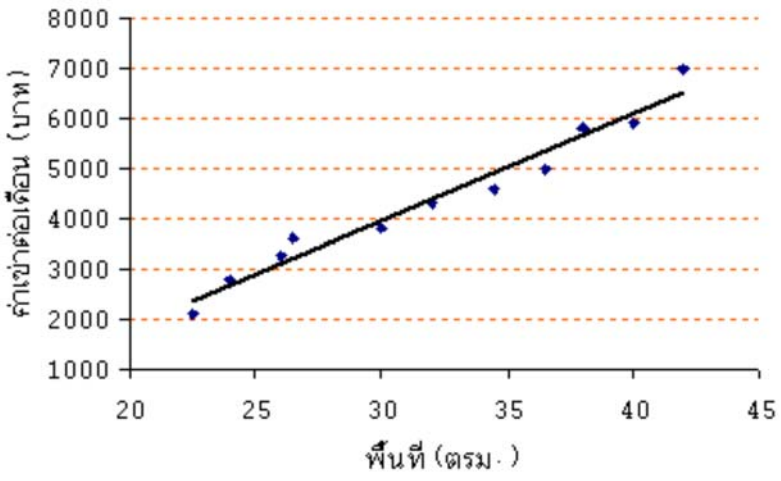
ตัวอย่างการประยุกต์ใช้ Simple Linear Regression Analysis

ผู้เขียนอยากจะยกตัวอย่างให้เห็นภาพวิธีการนำไปประยุกต์ใช้งาน ซึ่งจริงๆเวลาใช้งาน เรามักจะใช้ Software ช่วยในการคำนวณ ขั้นตอนก็จะรวบรัดมากกว่าที่ได้กล่าวในหัวข้อที่ผ่านมา นั่นเพราะว่าในการนำทฤษฎีอะไรก็ตามไปใช้งานจริงๆ เราก็ไม่จำเป็นต้องแสดงวิธีคิด หรือขั้นตอนอย่างละเอียด แต่นั่นต้องพึงระลึกอยู่เสมอว่า เราต้องมีความเข้าใจในทฤษฎี ซึ่งจะทำให้เราเข้าใจเหตุและผลของสิ่งที่ได้

ตัวอย่าง นักวิจัยการตลาดคนหนึ่งกำลังสำรวจข้อมูล ราคาเช่าต่อเดือนของอาคารชุดประเภทคอนโดมิเนี่ยม ในเขตใกล้เคียงกับถนนศรีนครินทร์ ในช่วงที่อยู่ในเขตกรุงเทพมหานคร โดยการสำรวจได้เก็บตัวอย่าง โดยใช้อัตราค่าเช่าห้องต่อเดือน และขนาดของห้อง เป็นเกณฑ์ โดยได้ข้อมูลมาดังตารางต่อไปนี้ (ตัวเลขสมมติ)

พื้นที่ (m^2)	ค่าเช่าต่อเดือน (บาท)
22.5	2100
24	2800
26	3250
26.5	3600
30	3800
32	4300
34.5	4600
36.5	5000
38	5800
40	5900
42	7000

เมื่อเราได้ข้อมูลมาดังในตาราง เราจะเริ่มการพิสูจน์ข้อมูลโดยใช้กราฟ Scatter plot



จากกราฟที่ได้พอที่จะบอกได้ว่า พื้นที่กับค่าเช่าต่อเดือนมีความสัมพันธ์กัน จริงๆแล้วในโลกแห่งความเป็นจริง ก็ต้องเป็นเช่นนั้น เพราะหากกราฟไม่ได้บอกว่า พื้นที่กับค่าเช่า ไม่มีความสัมพันธ์กันเลยนั่น จะเป็นเรื่องผิดปกติ แสดงว่าการได้ข้อมูลมาอาจจะผิด ผู้เขียนกำลังต้องการจะสื่อว่าคนที่กำลังจะวิเคราะห์ข้อมูลจำเป็นต้องมี Sense แห่งความเป็นจริง และต้องตรวจสอบหากเห็นความผิดปกติ ซึ่งอาจจะมาจากการบันทึกหรือใส่ข้อมูลผิดพลาด หรือไม่ก็เก็บข้อมูลไม่เป็นไปตามหลักการที่ถูกต้อง เช่น ไปเก็บข้อมูลของหอพักนักศึกษา มารวมกับข้อมูลจากคอนโดมิเนี่ยม หรือไปเก็บจากตัวอย่างแถวๆ ถนนสุขุมวิท บริเวณอโศก มาเทียบกับแถวศรีนครินทร์ ย่านสวนหลวง เพราะปัจจัยดังกล่าวมีผลต่อตัวแปรตาม (ค่าเช่าห้องต่อเดือน) อย่างมาก แม้พื้นที่จะไม่ต่างกันนักก็ตาม

ใช้ Software ทำการประมวลผล ในกรณีนี้ผู้เขียนขอใช้ Microsoft Excel เทียบกับ Minitab แล้วทำการสรุปผล เริ่มจาก Microsoft Excel

	Coefficients	Standard Error	t Stat	P-value
Intercept	-2487.472284	455.7671781	-5.45777	0.000402
X Variable 1	214.5232816	13.96587796	15.36053	9.18E-08

Regression Statistics	
Multiple R	0.981457
R Square	0.963257
Adjusted R Square	0.959175
Standard Error	296.59
Observations	11

เมื่อดูข้อมูลจาก 2 ตารางที่ได้จากโปรแกรม Microsoft Excel เราจะได้สมการ Regression ดังนี้

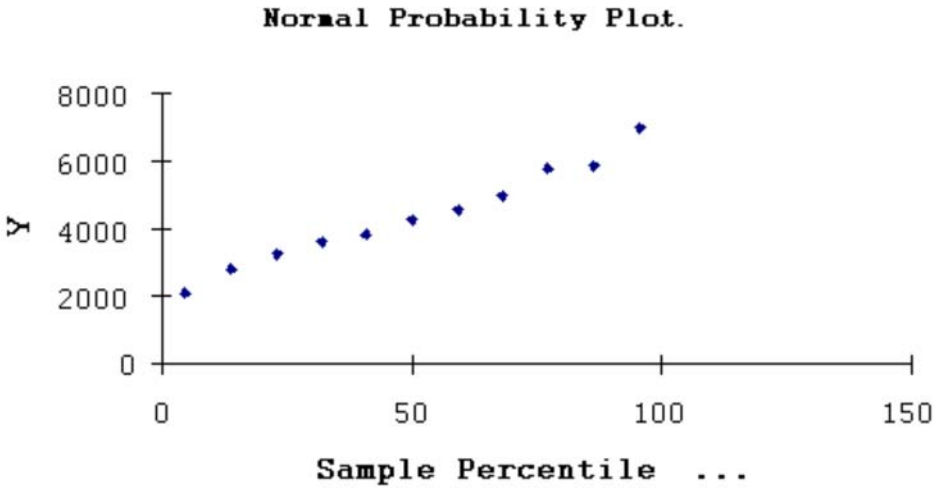
ค่าเช่าต่อเดือน = -2487.47 + 214.52 x พื้นที่ห้อง

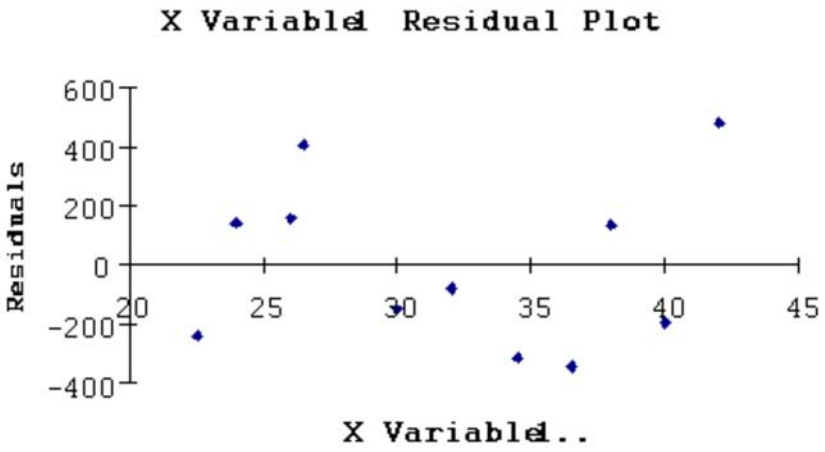
ทั้งนี้เพราะจากตารางที่ 1 ค่า P-Value ของค่าคงที่ และ สัมประสิทธิ์ของตัวแปรต้น(พื้นที่) มีค่าน้อยกว่า 0.05 ทั้งคู่ (กรณีนี้ใช้ค่าความเชื่อมั่นที่ 95%) ทำให้ไม่สามารถละทิ้งได้ เมื่อดูจากตารางที่ 2 ก็จะเห็นว่า ค่า R^2 และ R^2 Adjust ก็ผ่านเกณฑ์ทั้งคู่ จึงเชื่อได้ว่า ตัวแปรต้นและตัวแปรตาม ในการทดลองครั้งนี้มีความสัมพันธ์ต่อกัน แน่นอน แต่จะเป็นแบบเชิงเส้นหรือไม่ ต้องวิเคราะห์ต่อไปจาก ANOVA Table

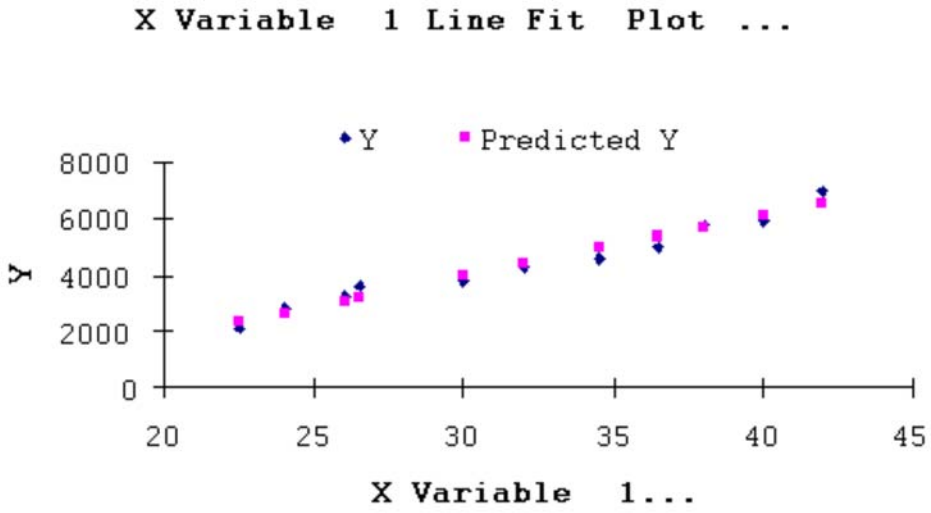
ANOVA					
	df	SS	MS	F	Significance F

Regression	1	20755127.49	20755127	235.9459	9.17675E-08
Residual	9	791690.6874	87965.63		
Total	10	21546818.18			

จากตาราง ANOVA ที่ได้ เราจะเห็นว่า Regression มีค่ามากกว่า Residual (Error) ถึง 235.9 เท่า นั่นแสดงว่าสาระส่วนใหญ่ที่ได้หากนำสมการนี้ไปใช้ ก็จะเป็น Regression มากกว่า Error หรือดูจากค่า P-Value ที่ถือได้ว่าเป็น 0 แสดงว่าสมการที่ได้มีความสัมพันธ์กันแบบเชิงเส้น (ปฏิเสธสมมติฐานที่ว่า สมการไม่ได้มีความสัมพันธ์กันแบบเชิงเส้น)







เมื่อเรา Verify ข้อมูล โดยใช้กราฟ 3 ข้อมูล ก็ยืนยันว่าข้อมูลที่เก็บมานั้นผ่านเกณฑ์หรือเงื่อนไขที่สำคัญ

ถ้าใช้ โปรแกรม Minitab จะได้ผลการวิเคราะห์ดังนี้

Regression Analysis: Rent versus Area

The regression equation is

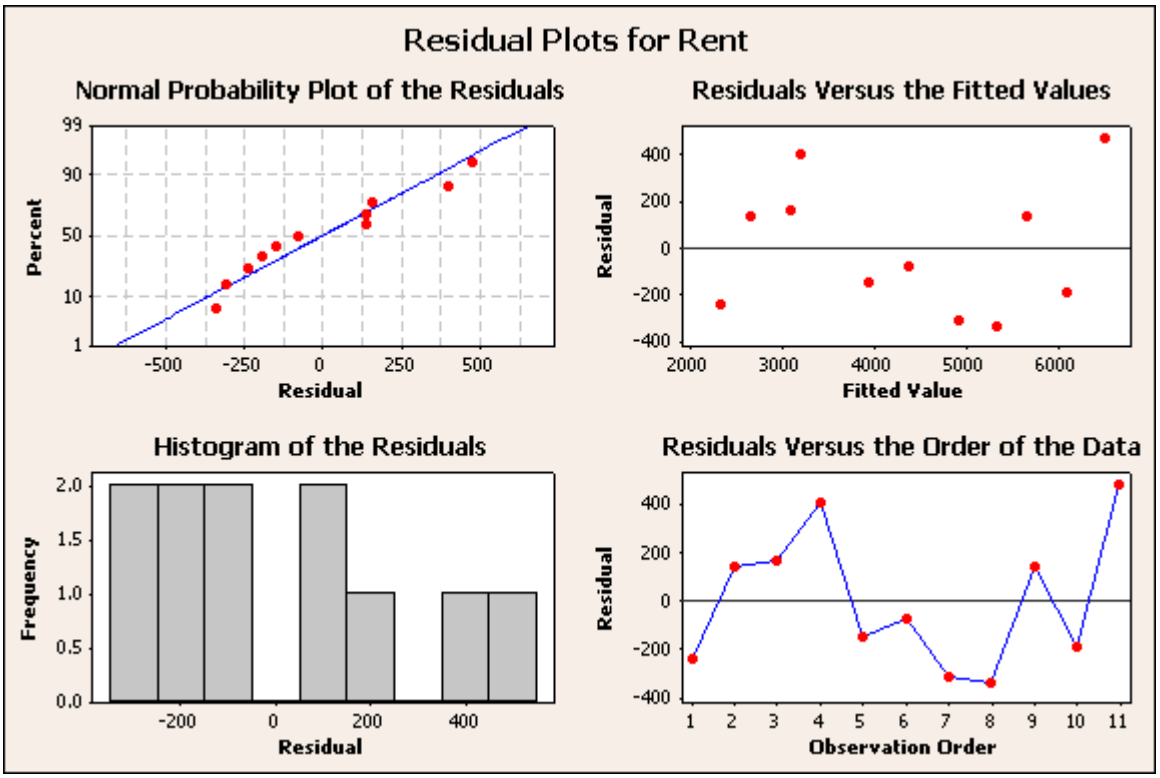
Rent = - 2487 + 215 Area

Predictor	Coef	SE Coef	T	P
Constant	-2487.5	455.8	-5.46	0.000
Area	214.52	13.97	15.36	0.000

S = 296.590 R-Sq = 96.3% R-Sq(adj) = 95.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	20755127	20755127	235.95	0.000
Residual Error	9	791691	87966		
Total	10	21546818			



มาถึงตรงนี้การวิเคราะห์ Regression ก็เสร็จสิ้น แล้วเราจะนำอะไรไปใช้ และใช้ได้อย่างไร เราจะนำสมการ Regression ที่ได้ไปใช้ประโยชน์ ยกตัวอย่าง พนักงานรายนี้จะใช้ข้อมูลดังกล่าวนำเสนอผู้บริหาร โดยนำสมการไปใช้คำนวณหาค่าเช่าที่จะกำหนด เพื่อให้ผู้บริหารตัดสินใจว่า ถ้าจะสร้างห้องขนาดเท่าที่ผู้บริหารกำหนดไว้นั้น ควรจะเก็บค่าเช่าเดือนละเท่าใด โดยเขาทำเป็นตารางดังตัวอย่างต่อไปนี้

พื้นที่ห้อง (ตารางเมตร)	ค่าเช่าต่อเดือน (บาท)
36	5,235.37
45	7,166.08
56	9,525.83

โดยใช้สมการ

ค่าเช่าต่อเดือน = -2487.47 + 214.52 x พื้นที่ห้อง

จะเห็นว่า เขาสามารถที่จะประมาณได้ว่า ห้องที่มีขนาดพื้นที่แตกต่างกัน ควรจะจะกำหนดราคาเช่าต่อเดือนเท่าใด ทั้งๆที่ข้อมูลในตารางนี้เขาไม่ได้เก็บข้อมูลมาก่อน แต่เมื่อเขามีสมการความสัมพันธ์ระหว่าง พื้นที่ห้องกับค่าเช่ารายเดือน เขาก็สามารถที่จะนำไปใช้คาดการณ์ได้

[HOME] [CONTENTS]

ขั้นตอนการวิเคราะห์ Multiple Linear Regression

ในหัวข้อนี้ จะศึกษาความสัมพันธ์ระหว่าง ตัวแปรตาม (Response , Dependent variable , Y) หนึ่งตัวกับตัวแปรอิสระ (Predictor, Independent variable, X) มากกว่าหนึ่งตัว แต่ความสัมพันธ์ดังกล่าวยังคงเป็นแบบเส้นตรงอยู่ ในชีวิตจริงแล้วจะมีน้อยมากที่ปัจจัยหนึ่งจะขึ้นอยู่กับปัจจัยหนึ่งเพียงอย่างเดียว ส่วนมากแล้วตัวแปรตามมักจะขึ้นอยู่กับตัวแปรอิสระหลายตัว พูดง่ายๆภาษานักสถิติคือ Y มักจะขึ้นอยู่ X หลายตัว นั่นเอง ดังตัวอย่างต่อไปนี้

ตัวอย่าง 1 ในการจะศึกษาประสิทธิภาพการใช้น้ำมันของรถยนต์ เราไม่สามารถจะเอาขนาดของเครื่องยนต์มาเป็นตัวกำหนดเพียงอย่างเดียว จะต้องคำนึงถึงน้ำหนักตัวรถ น้ำหนักคนขับ อายุของเครื่องยนต์ ความเสียดทานต่อผิวถนนของล้อรถ พูดง่ายๆคือหากต้องการพยากรณ์อัตราความสิ้นเปลืองของน้ำมันเชื้อเพลิง หรืออัตราการใช้น้ำมัน (กิโลเมตร/ลิตร) แล้วจะต้องคำนึงถึงตัวแปรอิสระมากกว่าหนึ่งตัวแปร

ตัวอย่างที่ 2 การจะศึกษาเพื่อพยากรณ์ปริมาณสารเคมีในกระแสเลือดของคนงานในโรงงานเคมี ตัวแปรอิสระที่ต้องใช้เป็นตัวคาดการณ์ จะประกอบด้วย

Y : ปริมาณสารเคมีในกระแสเลือด

X1 : จำนวนปีที่ทำงานอยู่กับสารเคมีนั้น

X2 : จำนวนปี (เดือน หรือ สัปดาห์) ที่ออกห่างมาจากสถานที่ทำงานแบบนั้น

X3 : อายุของคนงานนั้น

X4 : น้ำหนักของคนงาน หรือดัชนีอื่นที่บ่งบอกถึงมวลกาย

จะเห็น X ทั้งหมดล้วนเป็นปัจจัยที่จะทำให้สารเคมีเจือปนในกระแสเลือดมากน้อยได้ เช่น คนงานที่ทำงานมา 4 ปีกับ 1 ปี ย่อมได้รับสารเคมีในปริมาณที่ต่างกัน คนงานที่อายุ 22 ปีจะร่างกายจะยังมีความสามารถในการกำจัดสารแปลกปลอมในร่างกายได้ดีกว่าคนอายุ 35 ปี หรือคนที่ร่างกายใหญ่โตแข็งแรงก็จะมีขีดความสามารถในการจำกัดสิ่งแปลกปลอมในกระแสเลือดได้ดีกว่า คนตัวเล็ก คนผอม

ตัวอย่างที่ 3 การจะวัดความฟิตของนักกีฬาสามารถวัดผ่านปริมาตรออกซิเจนที่ร่างกายใช้ต่อนาทีได้ แต่การจะวัดให้ได้อย่างแม่นยำนั้นไม่ใช่เรื่องง่ายและยังสิ้นเปลืองค่าใช้จ่ายที่สูงมาก แต่เราสามารถวัดโดยวิธีอ้อมได้ดังต่อไปนี้

Y : ปริมาตรออกซิเจนในการหายใจ (ลิตร/นาที)

X1 : นำหนักของนักกีฬา (กิโลกรัม)

X2 : อายุของนักกีฬา (ปี)

X3 : ความสามารถในการเดิน โดยใช้ค่าเวลาที่เดินได้ 1 ไมล์ (นาที)

X4 : อัตราการเต้นของหัวใจเมื่อเดินได้ 1 ไมล์ (ครั้ง/นาที)

จากการวิจัยในสหรัฐอเมริกา ได้มีผลการศึกษาของมหาวิทยาลัยแห่งหนึ่งโดยทำการศึกษากับนักศึกษาจำนวนหนึ่ง โดยได้ตีพิมพ์ผลการศึกษาดังกล่าวในหัวข้อ " Validation of th Rockport Fitness Walking Test in College Male and Female " (Reserach Quarterly for Excercise and Sport, 1994 : 152-158) โดยได้มีการนำเสนอสมการความสัมพันธ์ระหว่างตัวแปรทั้งหลายดังนี้

$$y = 5.0 + 0.01X1 - 0.05X2 - 0.13X3 - 0.01X4$$

หากนำ Regression model ดังกล่าวไปคาดการณ์ ปริมาตรออกซิเจนที่นักศึกษาคนหนึ่งใช้ในการหายใจ โดยมีข้อมูลดังนี้ น้ำหนัก 76 กก. อายุ 20 ปี สามารถเดิน 1 ไมล์ได้โดยใช้เวลา 12 นาที และอัตราการเต้นของหัวใจเมื่อเดินได้ 1 ไมล์ดังกล่าวอยู่ที่ 140 ครั้งต่อนาที

$$y = 5.0 + 0.01(76) - 0.05(20) - 0.13(12) - 0.01(140)$$

$$y = 1.80 \text{ ลิตร/นาที}$$

ซึ่งการจะสรุปผล Fitness ของนักศึกษาคนนี้อยู่ในเกณฑ์ใด ก็นำค่า y = 1.80 ลิตร/นาที ไปเปรียบเทียบกับตารางมาตรฐานอีกทีหนึ่ง

ทั้ง 3 ตัวอย่างนั้น เป็นเพียงตัวอย่างง่ายๆ เพื่อชี้ให้ท่านผู้อ่านได้เข้าใจว่า Multiple Linear Regression คืออะไรและใช้อะไรได้บ้าง โดยทั่วไปในความสัมพันธ์นั้นจะมีตัวแปรต้นอยู่หลายตัว แต่ผู้เขียนขอยกตัวอย่างและแสดงขั้นตอนการวิเคราะห์กรณีที่มีตัวแปรต้นเพียงสองตัวโดยรูปแบบความสัมพันธ์หรือ Regression model จะเป็นดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

เมื่อ ε คือ Error ของ Model ซึ่งจะมีค่าเข้าหา 0 (ไม่มี Error) ซึ่งเราจะมองเป็น Normal distribution ที่อยู่รอบๆค่า 0 และมี variance อยู่ค่าหนึ่ง เมื่อเราใช้ Least square method จะได้สมการความสัมพันธ์ดังต่อไปนี้ (ผู้เขียนขอไม่กล่าวถึงที่ไปที่มาของสมการเหล่านี้)

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{i1} + \beta_2 \sum_{i=1}^{n} x_{i2} + ...+ \beta_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i1}^2 + \beta_2 \sum_{i=1}^{n} x_{i1}x_{i2} + ...+ \beta_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

$$\beta_0 \sum_{i=1}^{n} x_{i2} + \beta_1 \sum_{i=1}^{n} x_{i2}x_{i1} + \beta_2 \sum_{i=1}^{n} x_{i2}^2 + ...+ \beta_k \sum_{i=1}^{n} x_{i2}x_{ik} = \sum_{i=1}^{n} x_{i2}y_i$$

$$\beta_0 \sum_{i=1}^{n} x_{ik} + \beta_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \beta_k \sum_{i=1}^{n} x_{i1}x_{i2} + ...+ \beta_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i$$

ตัวอย่าง มีข้อมูลดังตารางที่ 1 เมื่อต้องการวิเคราะห์ Multiple linear regression มีขั้นตอนดังนี้

Observ. no.	1	2	3	4	5	6	7	8	9	10

Y	9.95	24.45	31.75	35.00	25.02	16.86	14.38	9.60	24.35	27.50
X_1	2	8	11	10	8	4	2	2	9	8
X_2	50	110	120	550	295	200	375	52	100	300

ตารางที่ 1

ขั้นตอนที่ 1 หาสมการที่จะใช้คำนวณ จากตารางที่ 1 มีตัวแปรต้นหรือตัวแปรอิสระ(X) อยู่สองตัว จำนวนข้อมูล (n) 10 ข้อมูล ดังนั้น Regression model จึงมีค่าคงที่และสัมประสิทธิ์ของตัวแปรอิสระที่ต้องหา คือ b_0,b_1 และ b_2 โดยสมการที่ใช้หาจึงเป็นดังต่อไปนี้

$$n\beta_0 + \beta_1\sum_{i=1}^{n}x_{i1} + \beta_2\sum_{i=1}^{n}x_{i2} = \sum_{i=1}^{n}y_i$$

$$\beta_0\sum_{i=1}^{n}x_{i1} + \beta_1\sum_{i=1}^{n}x_{i1}^2 + \beta_2\sum_{i=1}^{n}x_{i1}x_{i2} = \sum_{i=1}^{n}x_{i1}y_i$$

$$\beta_0\sum_{i=1}^{n}x_{i2} + \beta_1\sum_{i=1}^{n}x_{i2}x_{i1} + \beta_2\sum_{i=1}^{n}x_{i2}^2 = \sum_{i=1}^{n}x_{i2}y_i$$

ขั้นตอนที่ 2 คำนวณหาค่าเพื่อแทนลงในสมการ จากตารางที่ 1 เราจะทำการคำนวณค่าต่างๆตามสมการทั้งสาม โดยใช้ตาราง Excel ช่วยในการคำนวณ ซึ่งจะได้ดังตารางที่ 2

	Y	X_1	X_2	$(X_1)(X_2)$	$(X_1)^2$	$(X_2)^2$	$(Y)^2$	$(X_1)(Y)$	$(X_2)(Y)$
	9.95	2	50	100	4	2500	99	19.90	497.50
	24.45	8	110	880	64	12100	579.8	195.60	2,689.50
	31.75	11	120	1320	121	14400	1008.06	349.25	3,810.00
	35.00	10	550	5500	100	302500	1225	350.00	19,250.00
	25.02	8	295	2360	64	87025	626	200.16	7,380.90
	16.86	4	200	800	16	40000	284.26	67.44	3,372.00
	14.38	2	375	750	4	140625	206.78	28.76	5,392.50
	9.60	2	52	104	4	2704	92.16	19.20	499.20
	24.35	9	100	900	81	10000	592.92	219.15	2,435.00
	27.50	8	300	2400	64	90000	756.25	220.00	8,250.00
SUM	**218.86**	**64**	**2152**	**15114**	**522**	**701854**	**5488.24**	**1669.46**	**53576.6**

ตารางที่ 2

จากตารางที่ 2 เมื่อนำค่าที่ได้ใส่สมการทั้งสาม จะได้สมการใหม่ 3 สมการเรียงลำดับดังนี้

$$\beta_0(10)+\beta_1(64)+\beta_2(2152)=218.86$$
$$\beta_0(64)+\beta_1(522)+\beta_2(15114)=1669.46$$
$$\beta_0(2152)+\beta_1(15114)+\beta_2(701854)=53576.6$$

วิธีที่สามารถใช้ในการแก้สมการ เพื่อหาค่า b_0,b_1 และ b_2 นั้นมีอยู่หลายวิธี แต่ผู้เขียนขอเลือกใช้วิธี Matrix ในการแก้สมการเพื่อหาคำตอบ ซึ่งท่านสามารถอ่านวิธีการ Matrix เพื่อเป็นตัวอย่างได้ ทางนี้ <<< Link To Matrix >>>

ขั้นตอนที่ 3 เปลี่ยนสมการให้อยู่ในรูป Matrix แล้วใช้วิธีการทาง Matrix ในการหาค่า b_0,b_1 และ b_2 จะได้สมการในรูป

$$[Y]=[A][\beta]$$
$$\therefore \quad [\beta]=[A]^{-1}[Y]$$

เมื่อแทนค่าแล้วจะได้ Matrix เป็น

$$\begin{bmatrix}218.86\\1669.46\\53576.6\end{bmatrix}=\begin{bmatrix}10 & 64 & 2152\\64 & 522 & 15114\\2152 & 15114 & 701854\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\end{bmatrix}$$

เมื่อทำการหา Inverse matrix จะได้ Matrix เป็นดังต่อไปนี้ (ผู้อ่านควรทำความเข้าใจหลักการคูณกันของ Matrix ด้วย)

$$\begin{bmatrix}\beta_1\\\beta_2\\\beta_3\end{bmatrix}=\begin{bmatrix}0.5509466 & -0.0495021 & -0.0006233\\-0.0495021 & 0.0095360 & -0.0000536\\-0.0006233 & -0.0000536 & 0.0000045\end{bmatrix}\begin{bmatrix}218.86\\1669.46\\53576.6\end{bmatrix}$$

$$\beta_0=4.544$$
$$\beta_1=2.2158$$
$$\beta_2=0.0147$$

ดังนั้นสมการหรือ Regression model ที่ได้จะเป็น

$$Y=4.544+2.2158X_1+0.0147X_2$$

ขั้นตอนที่ 4 ทดสอบสมมติฐาน เพื่อหาว่าค่า b_0,b_1 และ b_2 ที่หามาได้นั้นมีนัยสำคัญความแตกต่างกับ 0 หรือไม่ พูดง่ายๆ คือจำเป็นต้องคงค่า หรือพจน์ที่ค่านี้คูณอยู่ไว้ใน Model หรือไม่ โดยสมมติฐานที่ต้องทดสอบคือ

$$H0:\beta_0=0 \quad | \quad H0:\beta_1=0 \quad | \quad H0:\beta_2=0$$
$$Ha:\beta_0\neq 0 \quad | \quad Ha:\beta_1\neq 0 \quad | \quad Ha:\beta_2\neq 0$$

ในการทดสอบ จะใช้ T -Statistics ทั้งนี้เพราะเรา Assume ว่า ค่า b_0, b_1 และ b_2จะเป็น Normal distribution รอบๆค่ากลางค่าหนึ่ง เรากำลังจะทดสอบว่าค่ากลางดังกล่าวเท่ากับ 0 หรือไม่ โดยสมการในการหาค่า T เป็นดังนี้ (ผู้เขียนขอไม่อธิบายที่มาของสมการ)

$$t_{\beta_i} = \frac{\beta_i}{\sqrt{(\sigma^2_{\beta_i})(C_{\beta_i})}}$$

$$\sigma^2_{\beta_i} = \frac{SS_{Error}}{n-p}$$

$$SS_{Error} = \sum_{i=1}^{n} Y_i^2 - [\beta][Y]$$

$$[\beta][Y] = [\beta_0 \quad \beta_1 \quad \beta_2] \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

เมื่อ

$(S_{bi})^2$: Estimated Standard Error

C_{bi} : ค่าที่ได้มาจาก Inverse matrix $[A]^{-1}$ ตามแนวทะแยงมุม ที่ตรงกับ b_i นั้นๆ

n : No of observation

p : No of regressor (b_0 , b_1 .... b_k) ตัวอย่างนี้ คือ 3

$$[\beta][Y] = [4.544 \quad 2.2158 \quad 0.0147] \begin{bmatrix} 218.86 \\ 1669.46 \\ 53576.6 \end{bmatrix}$$

$$= (4.544)(218.86) + (2.2158)(1669.46) + (0.0147)(53576.6)$$

$$= 5480.62$$

จากตารางที่ 2

$$\sum_{i=1}^{10} Y_i^2 = 5488.24$$

$$SS_{Error} = 5488.24 - 5480.62 = 7.624$$

$$\sigma^2_{\beta} = \frac{7.624}{10-3} = 1.089$$

จาก Inverse matrix ค่าตามแนวทะแยงลง จะได้

$$C_{\beta_0} = 0.5509466 \qquad C_{\beta_1} = 0.009536 \qquad C_{\beta_2} = 0.0000045$$

ดังนั้นจะได้

$$t_{\beta 0} = \frac{\beta_0}{\sqrt{(\sigma_\beta^2)(C_{\beta_0})}} = \frac{4.544}{\sqrt{(1.089)(0.5509466)}} = 5.8665$$

$$t_{\beta 1} = \frac{\beta_1}{\sqrt{(\sigma_\beta^2)(C_{\beta_1})}} = \frac{2.2158}{\sqrt{(1.089)(0.009536)}} = 21.7485$$

$$t_{\beta 2} = \frac{\beta_2}{\sqrt{(\sigma_\beta^2)(C_{\beta_2})}} = \frac{0.0147}{\sqrt{(1.089)(0.0000045)}} = 6.6476$$

ถ้ากำหนด **a** = 0.05 เมื่อเปิดตาราง T เพื่อหาค่าวิกฤติ จะได้

$$t_\beta - Critical = t_{\alpha/2,\, df=n-p} = t_{0.025,7} = 2.365$$

เราจะปฏิเสธสมมติฐานหลัก ถ้าค่า t_b ที่คำนวณได้มากกว่า t_b วิกฤติที่หาได้จากตาราง T ดังนั้นสมมติฐานหลักทั้งสามจึงไม่เป็นจริง นั่นคือ ค่า b_0, b_1 และ b_2 มีค่าไม่เท่ากับ 0 จริง จึงไม่สามารถตัดออกจาก Regression model ได้

ขั้นตอนที่ 5 การพิสูจน์ว่า Regression model ที่ได้มานั้นเหมาะที่จะนำไปใช้คาดการณ์ (Predict) ค่า Y ในอนคตมากน้อยเพียงใด ซึ่งจะใช้วิธีพิสูจน์ค่าความคลาดเคลื่อน (Error) ตัวสถิติที่จะใช้ทดสอบความคลาดเคลื่อนนี้ เราเรียกว่า F-Statistic และสมมติฐานคือ

H_0 : Error จากการใช้ Model นี้ Predict ค่า Y เป็น Error ที่ไม่สามารถอธิบายได้เป็นส่วนใหญ่

H_a: Error จากการใช้ Model นี้ Predict ค่า Y เป็น Error ที่สามารถอธิบายได้เป็นส่วนใหญ่

สมการทางคณิตศาสตร์ที่ใช้ในการคำนวณ มีดังนี้

$$F = \frac{MS_{Regression}}{MS_{Error}}$$

$$MS_{Regression} = \frac{SS_{Regression}}{k}$$

$$MS_{Error} = \frac{SS_{Error}}{n - p}$$

ค่า degree of freedom หาได้จาก

Total = n-1 = 10-1 = 9

Error = n - p = 10 - 3 = 7

Regression = k = Total - Error = 9 - 7 = 2

$$SS_{Regression} = [\beta][Y] - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$= \begin{bmatrix} 4.544 & 2.2158 & 0.0147 \end{bmatrix} \begin{bmatrix} 218.86 \\ 1669.46 \\ 53576.6 \end{bmatrix} - \frac{(218.86)^2}{10}$$

$$= 5480.62 - 4790$$

$$= 690.62$$

$$MS_{Regression} = \frac{690.62}{2} = 345.31$$

$$MS_{Error} = \frac{7.624}{7} = 1.089$$

$$\therefore \quad F = \frac{345.31}{1.089} = 317.047$$

หาค่า F-critical จากตาราง F

$$F\text{-}critical = F_{\alpha,v1,v2} = F_{0.05,2,7} = 4.74$$

สมมติฐานหลักจะไม่เป็นจริง ถ้าค่า F ที่คำนวณได้ มากกว่า F-critical ที่ได้จากตาราง ดังนั้นกรณีนี้เราจึงปฏิเสธสมมติฐานหลัก นั่นคือ Error ของ Model นี้ส่วนใหญ่สามารถอธิบายได้ (เกิดจากการเปลี่ยนค่า X1 หรือ X2) มากกว่าจะเกิดจากเหตุอื่นๆ จึงสรุปว่า Regression model นี้ ให้ความแม่นยำสูงถ้านำไปพยากรณ์ค่า Y

ขั้นตอนที่ 6 การหา Coefficient of Determination

$$R^2 = \frac{(SS_{Regression})100\%}{SS_{Total}}$$
$$= \frac{(SS_{Regression})100\%}{SS_{Regression} + SS_{Error}}$$
$$= \frac{(690.62)100\%}{690.62 + 7.694}$$
$$= 98.9\%$$

$$R^2 - Adjusted = \left\{1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)\right\}(100\%)$$
$$= \left\{1 - \left(\frac{10-1}{10-3}\right)(1 - 0.989)\right\}(100\%)$$
$$= 98.5\%$$

พบว่า ค่า R^2 มีค่าสูงมาก R^2-adjusted ก็ต่ำกว่า R^2 ไม่มาก สรุปว่า Error ที่เราไม่สามารถอธิบายได้มีมากกว่า Error ที่เราไม่สามารถอธิบายที่มาได้ ในอัตราส่วนที่มากทีเดียว และจำนวนสิ่งตัวอย่างที่เก็บมานั้นก็อยู่ในเกณฑ์มาตรฐาน

หากนำค่าที่ได้จากการคำนวณมาเขียนสรุปเป็นตารางจะได้ดังต่อไปนี้

b_0	$t = 5.8665$	F = 317.074
b_1	$t = 21.7485$	
b_2	$t = 6.6476$	
R^2	0.989	
R^2-Adjusted	0.985	

ตารางที่ 3

ในกรณีที่เราใช้โปรแกรม Microsoft Excel ช่วยในการวิเคราะห์ จะได้ตารางออกมาดังต่อไปนี้

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9945
R Square	0.9891
Adjusted R Square	0.9860
Standard Error	1.0436
Observations	10

ANOVA					
	df	SS	MS	F	Significance F

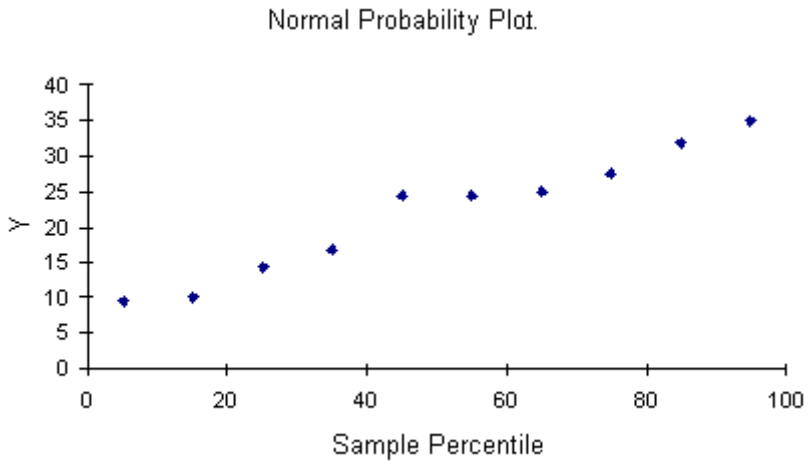
Regression	2	690.6502	345.3251	317.0505	0.0000
Residual	7	7.6243	1.0892		
Total	9	698.2744			

	Coefficients	Standard Error	t Stat	P-value
Intercept	4.5444	0.7746	5.8663	0.0006
X1	2.2158	0.1019	21.7422	0.0000
X2	0.0147	0.0022	6.6410	0.0003

ตารางที่ 4

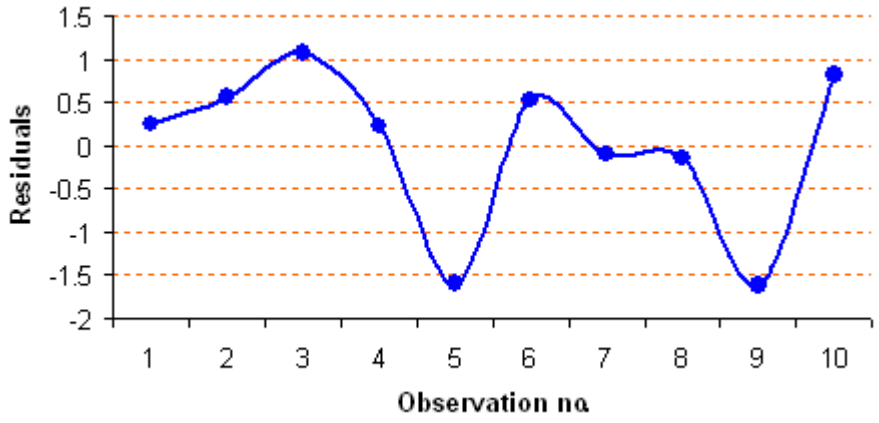
ขั้นตอนที่ 7 การพิสูจน์คุณสมบัติ 3 ประการ โดยกราฟที่ได้จากโปรแกรม Excel

- Normality



จากกราฟ การเรียงตัวของจุดค่า Y เทียบกับ Percentile เป็นแนว แม้จะไม่เป็นเส้นตรงเสียทีเดียว แต่สามารถยอมรับได้ว่าเป็น Normal distribution ได้

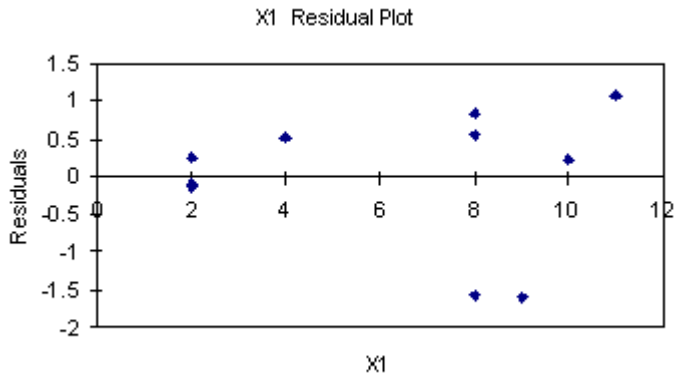
- Independence

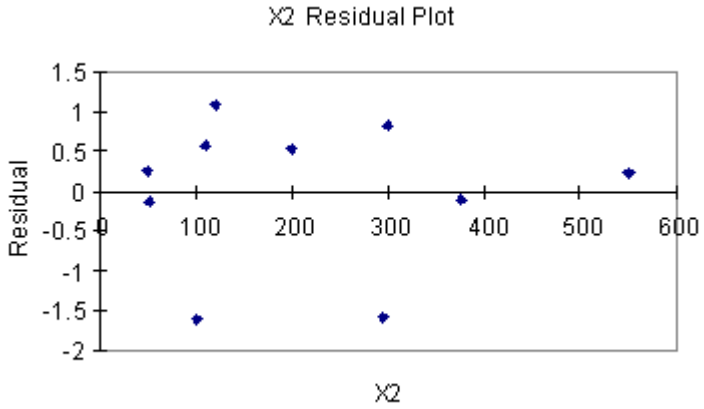


จะพบว่า แนวของจุดถือได้ว่า ไม่มีทิศทางใดแน่นอน ไม่ได้อยู่ทางด้านลบหรือบวกอย่างเดียว ไม่ได้ขึ้นหรือลงอย่างเดียว ลักษณะเช่นนี้เราถือว่าความเป็นอิสระของ X แต่ละตัวอยู่ในเกณฑ์ที่ยอมรับได้ (กรณี Multiregression Analysis โปรแกรม

Excel ไม่ได้พล้อตให้ ท่านจำเป็นต้องทำเอง)

- Homoscedasticity





เมื่อทำการพล้อต Residual กับค่า X (Fit) ทั้งสอง (X) พบว่าจุดไม่มีลักษณะอยู่ด้านบวก หรือลบตลอด หรือเป็น 0 ตลอด หรือกว้างออกตลอด เมื่อค่า X สูงขึ้นหรือต่ำลง เราพอจะอนุมานได้ว่า Residual ตลอดย่านค่า X ไม่ได้แตกต่างกันจนเกินเหตุ นั่นคือการเพิ่มหรือลดค่า X ไม่ได้ทำให้ความคลาดเคลื่อนหรือ Error ของ Regression model เปลี่ยนไปจนเกินเหตุ เราจะถือว่าผ่านเงื่อนไขนี้ (แยกวิเคราะห์แต่ละ X)

จะเห็นว่า แม้จะมีตัวแปรต้น(อิสระ) หรือ X มากกว่าหนึ่งตัว แต่เราก็ยังใช้วิธีวิเคราะห์เหมือนกัน แตกต่างกันเฉพาะรายละเอียดเท่านั้น

[HOME] [CONTENTS]

Matrix

Matrix เป็นเทคนิคทางคณิตศาสตร์ที่ใช้ในการหาคำตอบหรือหาค่าตัวแปรต้นในสมการเชิงเส้นที่มีมากกว่า 1 สมการ โดยจะได้คำตอบหรือค่าตัวแปรที่หาพร้อมกัน ขั้นตอนการคำนวณจะเริ่มจากการจัดรูป Matrices ให้อยู่ในรูปทางคณิตศาสตร์

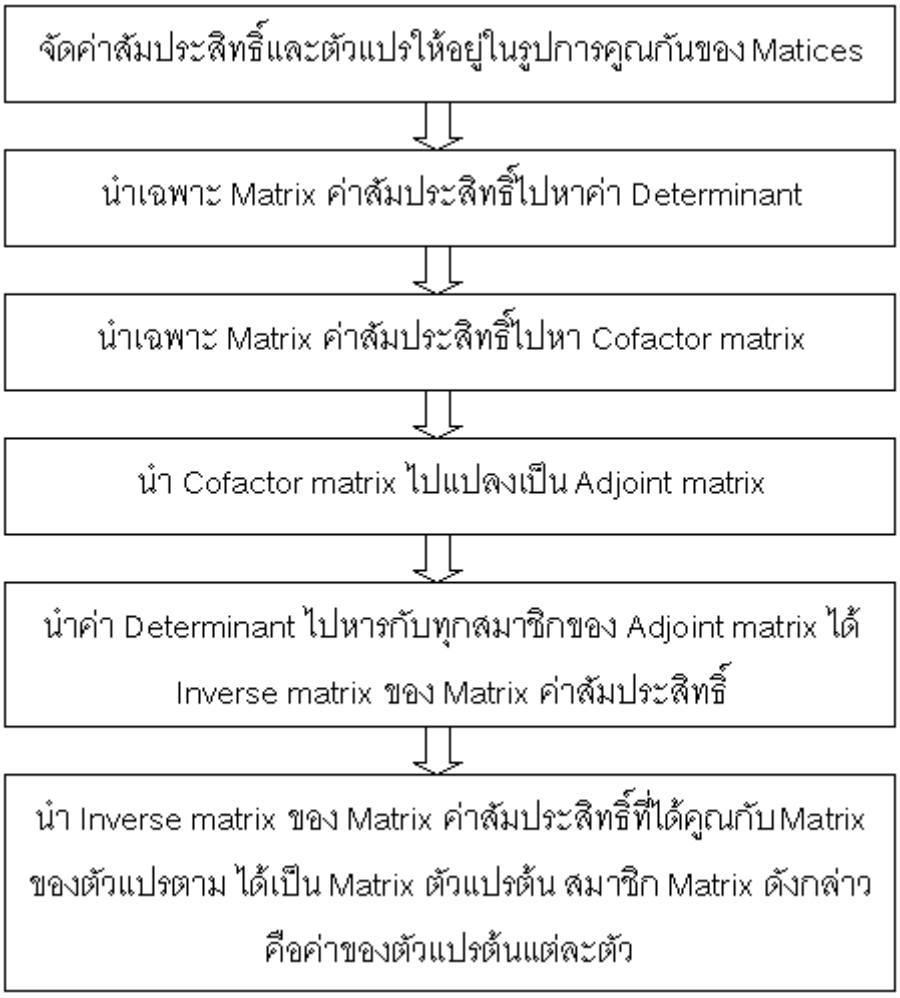
$$[Y] = [A][X]$$

โดย [Y] , [A] และ [X] คือ Matrix ที่เราต้องการหาคือ [X] จึงต้องจัดรูปสมการใหม่เป็นดังนี้

$$[X] = [A]^{-1}[Y]$$

เมื่อ $[A]^{-1}$ คือ Inverse matrix ของ [A] เนื่องจากในการคำนวณทางคณิตศาสตร์ Matrix จะไม่มีการหาร จึงต้องใช้วิธีหา Inverse matrix ของ [A] แล้วนำไปคูณกับ [Y] เพื่อให้ได้ [X] :ซึ่งทำให้การหาค่าตัวแปรในสมการเชิงเส้น โดยใช้ Matrix มีไม่ได้ทำง่ายๆเหมือนการแก้สมการแบบธรรมดาทั่วไป แต่ถ้ามีสมการเชิงเส้นหลายๆสมการ การใช้ Matrix ถือว่าง่ายกว่าการแก้สมการแบบธรรมหลายเท่านัก

มีขั้นตอนการคำนวณตามแผนภาพต่อไปนี้



ตัวอย่าง จงหาค่า X_1,X_2 และ X_3 จากสมการเชิงเส้นต่อไปนี้

$$15 = 2x_1 + 4x_2 + 2x_3$$
$$8 = 3x_1 + x_2 + x_3$$
$$6 = x_1 + x_3$$

จากสมการ จัดให้อยู่ในรูปการคูณกันของ Matrices จะได้ดังนี้

$$\begin{bmatrix} 15 \\ 8 \\ 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 2 \\ 3 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

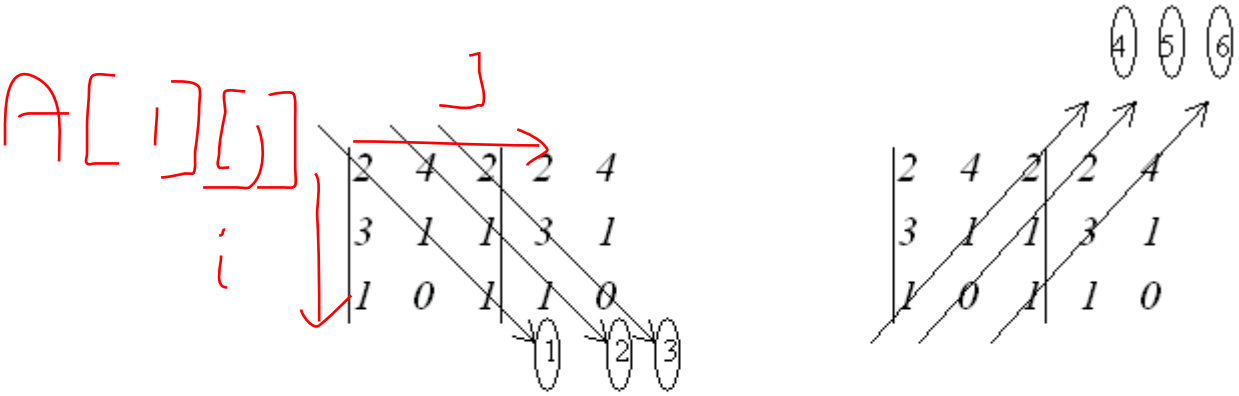
นำเฉพาะ Matrix สัมประสิทธิ์ของตัวแปรต้น ไปหาค่า Determinant คือค่าคงที่ที่แทนคุณสมบัติเฉพาะของ Matrix นั้นๆ

$$Det = \begin{vmatrix} 2 & 4 & 2 \\ 3 & 1 & 1 \\ 1 & 0 & 1 \end{vmatrix}$$

ขั้นแรก เอาค่าที่อยู่ใน Column 1,2 ไปต่อท้าย Matrix จะได้ Matrix ใหม่หน้าตาแบบนี้

$$\begin{vmatrix} 2 & 4 & 2 & 2 & 4 \\ 3 & 1 & 1 & 3 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{vmatrix}$$

ขั้นต่อไป นำค่าที่อยู่ในแนวเส้นทะแยงลงคูณกันแล้วนำมารวมกัน ลบด้วยผลรวมของค่าตามแนวทะแยงขึ้นคูณกัน ค่าที่ได้คือค่า Determinant



{(2x1x1)+(4x1x1)+(2x3x0)} - {(1x1x2)+(0x1x2)+(1x3x4)}

{2+4} - {2+12}

- 8

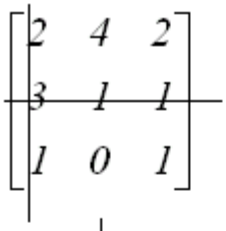
Determinant = -8

นำ Matrix สัมประสิทธิ์ไปหา Cofactor matrix วิธีการคือหาค่า Determinant ย่อยของ Matrix ส่วนที่เหลือเมื่อทำการตัด Row และ Column ของตัวสมาชิกที่เราจะหา Cofactor แล้วนำค่า Determinant ย่อยที่ได้ทั้งหมดมาเขียนอยู่ในรูป Matrix แล้วนำไปคูณกับ Matrix เครื่องหมายเอกลักษณ์ ผลลัพธ์ที่ได้จะเป็น Matrix ที่ค่าสมาชิกจะมีเครื่องหมายเปลี่ยนแปลงไปซึ่งแล้วแต่เครื่องหมายเดิมของสมาชิกนั้นๆกับเครื่องหมายของสมาชิกใน Matrix เครื่องหมายเอกลักษณ์ เราจะเรียกสิ่งที่ได้ว่า Cofactor matrix นั่นเอง

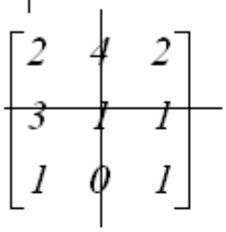
$$\begin{bmatrix} 2 & 4 & 2 \\ 3 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$ C11 = (1x1)-(0x1) = 1

$$\begin{bmatrix} 2 & 4 & 2 \\ 3 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$ C12 = (3x1)-(1x1) = 2

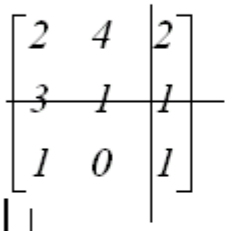
$$\begin{bmatrix} 2 & 4 & 2 \\ 3 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$ C13 = (3x0)-(1x1) = -1



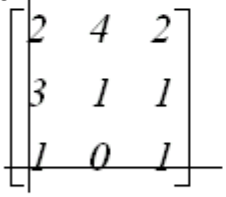
C21 = (4x1)-(0x2) = 4



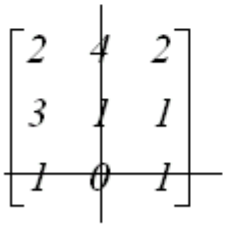
C22 = (2x1)-(1x2) = 0



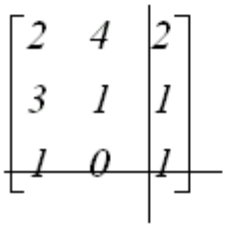
C23 = (2x0)-(1x4) = - 4



C31 = (4x1)-(1x2) = 2



C32 = (2x1)-(3x2) = - 4



C33 = (2x1)-(3x4) = -10

เมื่อนำมาเขียน Matrix จะได้

$$\begin{bmatrix} 1 & 2 & -1 \\ 4 & 0 & -4 \\ 2 & -4 & -10 \end{bmatrix}$$

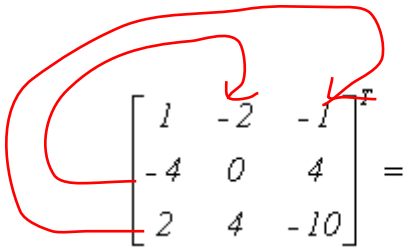
Matrix เครื่องหมายเอกลักษณ์ ของ Matrix ขนาด 3x3 คือ

$$\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$$

เอาเครื่องหมายคูณเข้า ในลักษณะ ตำแหน่งเดียวกันคูณกันจะได้ Cofactor matrix ดังนี้

$$\begin{bmatrix} 1 & -2 & -1 \\ -4 & 0 & 4 \\ 2 & 4 & -10 \end{bmatrix}$$

หา Adjoint matrix โดยการนำ Cofactor matrix ไปทำ Transpose (กลับตำแหน่งตาม Row ไปเป็นตาม Column)



$$\begin{bmatrix} 1 & -2 & -1 \\ -4 & 0 & 4 \\ 2 & 4 & -10 \end{bmatrix}^{T} = \begin{bmatrix} 1 & -4 & 2 \\ -2 & 0 & 4 \\ -1 & 4 & -10 \end{bmatrix}$$

หา Inverse matrix โดยการนำค่า Determinant มาหารค่าสมาชิกแต่ละค่า

$$Inverse = \frac{-1}{8}\begin{bmatrix} 1 & -4 & 2 \\ -2 & 0 & 4 \\ -1 & 4 & -10 \end{bmatrix} = \begin{bmatrix} -0.125 & 0.5 & -0.25 \\ 0.25 & 0 & -0.5 \\ 0.125 & -0.5 & 1.25 \end{bmatrix}$$

หาค่าตัวแปรต้นทั้งหมดได้โดยการเอา Matrix ตัวแปรตาม ([Y]) คูณกับ Inverse matrix ที่หาได้

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -0.125 & 0.5 & -0.25 \\ 0.25 & 0 & -0.5 \\ 0.125 & -0.5 & 1.25 \end{bmatrix}\begin{bmatrix} 15 \\ 8 \\ 6 \end{bmatrix}$$

$$= \begin{bmatrix} (-0.125)(15)+(0.5)(8)+(-0.25)(6) \\ (0.25)(15)+(0)(8)+(-0.5)(6) \\ (0.125)(15)+(-0.5)(8)+(1.25)(6) \end{bmatrix}$$

$$= \begin{bmatrix} 0.625 \\ 0.75 \\ 5.375 \end{bmatrix}$$

$$\therefore \quad x_1 = 0.625 \quad x_2 = 0.75 \quad x_3 = 5.375$$

[HOME] [CONTENTS]

Collinearity

คือสภาพที่เกิดสหสัมพันธ์ (Correlation) กันเองระหว่างตัวแปรอิสระในระดับค่อนข้างสูง เมื่อทำการวิเคราะห์ Multiple linear regressions เพื่อให้ท่านผู้อ่านเห็นภาพ ผู้เขียนขอยกตัวอย่างการเก็บข้อมูลเพื่อทำการศึกษาในเรื่องอัตราการเสียชีวิตของทารกตั้งแต่แรกเกิดไปถึงระยะเวลา 3 สัปดาห์หลังคลอด โดยมีการระบุตัวแปรอิสระใน Regression model ดังต่อไปนี้

อัตราการเสียชีวิตหลังคลอดของทารก ขึ้นอยู่กับระยะเวลาตั้งครรภ์ (สัปดาห์) และ น้ำหนักทารกแรกเกิด(กก.)

ท่านลองสังเกตดูดีๆจะพบว่าในความเป็นจริงแล้ว ระยะเวลาที่อยู่ในครรภ์มารดาก่อนคลอดของทารกที่สิ้นเกินไป นอกจากจะเป็นสาเหตุที่ทำให้อัตราการเสียชีวิตหลังคลอดสูงแล้ว ยังเป็นสาเหตุที่ทำให้ทารกแรกเกิดมีน้ำหนักน้อยกว่ามาตรฐานอีกด้วย ครั้นผู้ทำการวิจัยจะสรุปว่าน้ำหนักของทารกเป็นหนึ่งในสาเหตุของการเสียชีวิตของทารกหลังคลอดก็สรุปไม่ได้เต็มปากนัก เพราะระยะเวลาที่ทารกอยู่ในครรภ์ก็เป็นเหตุให้น้ำหนักทารกต่ำกว่าเกณฑ์มาตรฐานด้วย ครั้นจะเลือกเอาตัวแปร ระยะเวลาที่ทารกอยู่ในครรภ์ เป็นตัวแปรอิสระเพียงตัวเดียว ในความเป็นจริงก็จะมีกรณีที่อายุครรภ์ได้ตามเกณฑ์มาตรฐาน แต่น้ำหนักทารกไม่ได้มาตรฐานก็มี หรืออายุครรภ์น้อยกว่าเกณฑ์มาตรฐานแต่น้ำหนักได้มาตรฐานก็มีเช่นกัน ถ้าสมมติผู้ทำการวิจัยเลือกใช้ Model ที่มีตัวแปรอิสระสองตัวอย่างที่ว่านี้ ก็จะเกิดสภาพที่เรียกว่ามีสหสัมพันธ์กัน ระหว่างตัวแปร ระยะเวลาตั้งครรภ์ (สัปดาห์) และ น้ำหนักทารกแรกเกิด (กก.) ค่อนข้างสูง ที่เราเรียกว่า Collinearity

หรืออีกตัวอย่างหนึ่ง เรารู้ว่าอัตราสิ้นเปลืองน้ำมันของรถยนต์ขึ้นอยู่กับตัวแปรอิสระหลายตัวคือ ขนาดของเครื่องยนต์ ความเร็วที่ใช้ขับขี่ น้ำหนักบรรทุกและสัมประสิทธ์ความเสียดทานระหว่างยางล้อรถกับผิวถนน (เป็นต้น) เราพบว่ายิ่งน้ำหนักบรรทุกมากขึ้นเท่าใด ค่าสัมประสิทธ์ความเสียดทานระหว่างยางล้อรถกับผิวถนนก็จะมากขึ้น ลักษณะเช่นนี้คือการมีสหสัมพันธ์ระหว่างตัวแปรน้ำหนักบรรทุกและสัมประสิทธ์ความเสียดทานระหว่างยางล้อรถกับผิวถนน แล้วยังไปเกิดสหสัมพันธ์กับความเร็วรถยนต์อีกด้วย วนเวียนกันหลายความสัมพันธ์ ลักษณะเช่นนี้จะเรียกว่า Multicollinearity คือมีสหสัมพันธ์กันเองระหว่างตัวแปรอิสระมากกว่า 2 ตัวขึ้นไป นั่นเอง

Collinearity หรือ Multicollinearity ถึงแม้จะไม่ได้ทำให้ Model นั้นใช้ Predict ตัวแปรตามไม่ได้เลยก็ตาม แต่ปัญหาจะเกิดที่การจะควบคุมตัวแปรอิสระให้เป็นไปตาม Model จะไม่ใช่เรื่องง่ายอีกต่อไป ลักษณะเช่นนี้เราเรียกว่ามีปัญหา Reliability ของ Model คือลักษณะที่ใช้พยากรณ์แล้วจะได้ค่าตัวแปรตาม ไม่เหมือนเดิมตลอดเวลา ขึ้นอยู่กับสภาวะของตัวแปรอิสระที่มีสหสัมพันธ์กันด้วย เพราะนอกจากตัวแปรตามจะเปลี่ยนแปลงตามตัวแปรอิสระที่เปลี่ยนไปแล้ว ตัวแปรอิสระบางตัวยังเปลี่ยนแปลงโดยขึ้นอยู่กับตัวแปรอิสระตัวอื่นๆ อีกขั้น เลยเกิดความไม่มีเสถียรภาพของ Model ในต่างเวลากัน

ตัวอย่าง ถ้ามีข้อมูลดังในตารางและจะวิเคราะห์โดยใช้ Multiple regression ให้พิจารณา Multicollinearity

Y	X1	X2	X3	X4
125	13	18	25	11
158	39	18	39	30
207	52	50	62	43
182	42	43	50	29

196	50	37	65	46
175	44	29	59	32
145	11	27	24	14
144	22	23	31	17
160	30	18	34	22
175	51	31	58	30
151	27	25	29	21
161	41	22	53	22
200	51	52	75	36
173	37	36	44	27
175	43	38	37	20
162	43	28	45	16
155	38	19	40	18
230	62	56	75	50
162	28	30	36	20
153	30	25	41	23

ตารางที่ 1 ข้อมูลที่บันทึกไว้ก่อนทำการวิเคราะห์

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.987
R Square	0.974
Adjusted R Square	0.968
Standard Error	4.420
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	4	11127.940	2781.985	142.418	0.000
Residual	15	293.010	19.534		
Total	19	11420.950			

	Coefficients	Standard Error	t Stat	P-value
Intercept	99.513	3.313	30.037	0.0000
X Variable 1	0.677	0.183	3.706	0.0021
X Variable 2	0.925	0.138	6.722	0.0000
X Variable 3	-0.147	0.188	-0.781	0.4471
X Variable 4	0.845	0.202	4.192	0.0008

ตารางที่ 2 ผลการวิเคราะห์ Multiple regression โดยโปรแกรม Excel

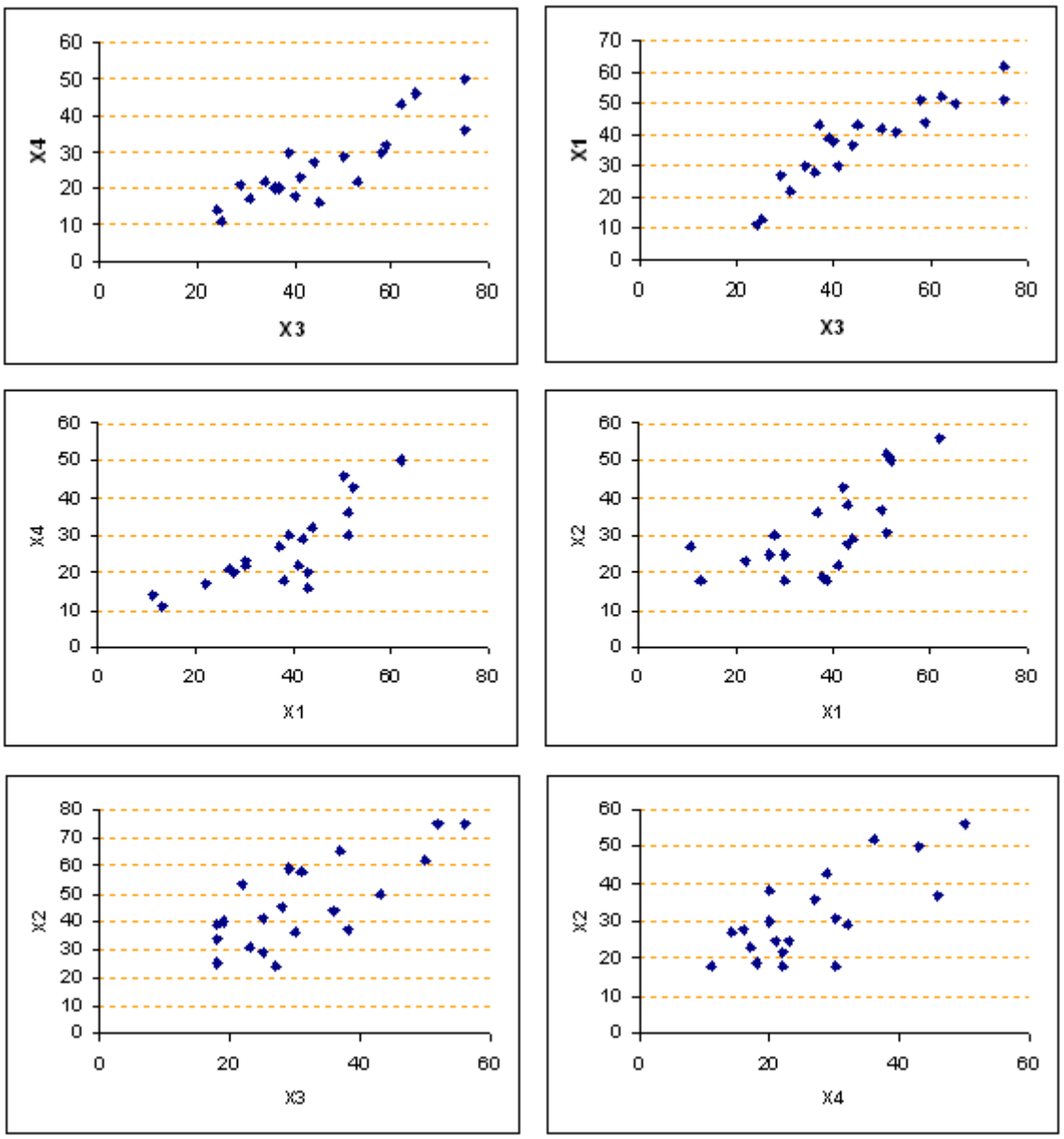
จะรู้ได้อย่างไรว่าเกิด Collinearity หรือ Multicollinearity ขึ้นแล้วเมื่อเราทำการวิเคราะห์ข้อมูลโดย Multiple regression

วิธีที่ 1 ง่ายที่สุดคือดูจากค่า F- Significane ของ Model (Regression) จากตาราง ANOVA และค่าทดสอบทาง

สถิติของสัมประสิทธิ์ตัวแปรอิสระแต่ละตัว โดยที่หาก F-Significane น้อยกว่า a (0.05) แปลว่า Regression model ดังกล่าวมีค่านัยสำคัญ แต่ถ้าค่าทดสอบทางสถิติของสัมประสิทธิ์ตัวแปรอิสระทั้งหมด หรือบางตัวไม่มีนัยสำคัญ (P-Value มากกว่า a) แปลว่ามีโอกาสเกิด Collinearity ระหว่างตัวแปรอิสระอย่างมากที่เดียว จากตารางที่ 2 เมื่อวิเคราะห์ Multiple linear regression โดยโปรแกรม Excel ตามขั้นตอนปกติ จะพบว่าค่า F-Significance บ่งบอกว่า Regression model มีนัยสำคัญ แต่เมื่อดู P-Value ของ X3 บ่งบอกว่า X3 ไม่มีนัยสำคัญต่อ Regression model เลยหรือบอกว่า ไม่จำเป็นต้องมี X3 เลยก็ได้ ลักษณะเช่นนี้คือมีโอกาสเกิด Collinearity สูงมาก

จริงหรือที่ไม่จำเป็นต้องมี X3 ใน Model ที่ได้

วิธีที่ 2 ใช้ Scatter plot ระหว่างตัวแปรอิสระทุกกคู่ จากรูปที่ 1 จะพบว่า คู่ X3,X4 คู่ X1,X3 และคู่ X1,X4 มีความสัมพันธ์กันอย่างมากที่เดียว ในขณะที่คู่อื่นๆที่เหลือก็มีความสัมพันธ์เชิงเส้นต่อกันเองพอสมควรที่เดียว โดยดูจากแนวการเรียงตัวของจุด กราฟที่ได้บ่งบอกว่าเกิด Multicollinearity ขึ้นแล้ว



รูปที่ 1 ตัวอย่าง Scatter plot ระหว่างตัวแปรอิสระ 6 คู่

วิธีที่ 3 ทดสอบหาค่าสหสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามและกับตัวแปรอิสระตัวอื่นๆ

	Y	X1	X2	X3
X1	0.887			
X2	0.896	0.687		
X3	0.892	0.905	0.754	
X4	0.916	0.819	0.744	0.871

ตารางที่ 3 Matrix ค่าสหสัมพันธ์ของทั้งตัวแปรตามและตัวแปรอิสระ

เมื่อใช้โปรแกรมคอมพิวเตอร์วิเคราะห์จะได้ค่าสหสัมพันธ์ (Pearson correlation : r) ดังตารางที่ 3 ตัวแปรอิสระทุกตัวมีความสัมพันธ์กับตัวแปรตาม โดยดูได้จากที่ค่าต่ำที่สุดก็ 0.887 (Y กับ X1) แล้ว ในขณะที่ ค่าสหสัมพันธ์ระหว่างคู่ตัวแปรอิสระเองก็ มีค่ามากตั้งแต่ 0.687 (X1 กับ X2) ขึ้นไปเลยทีเดียว ซึ่งถือว่าสูงมาก ยืนยันได้ว่าเกิด Multicollinearity ใน Regression model นี้ ทั้งๆที่จากตารางที่ 1 เราพบว่า X3 ไม่มีนัยสำคัญ แต่ค่าจากตารางที่ 3 บ่งบอกว่า X3 มีความสัมพันธ์กับ Y ในระดับที่สูงมาก

ค่าเท่าใดถึงจะถือว่ามี Collinearity โดยทั่วไปเราจะเปรียบเทียบค่าสหสัมพันธ์ระหว่าง X นั้นๆ กับ Y ถ้าน้อยกว่าเมื่อเทียบกับค่าสหสัมพันธ์กับ X ตัวอื่นๆ แสดงว่ามีโอกาสเกิด Collinearity สูง

วิธีที่ 4 วัดระดับ Multicollinearity ด้วยค่า Variance Inflation Factor (VIF)

เริ่มต้นเราพิจารณาค่า Variance ของค่าสัมประสิทธิ์แต่ละตัวแปรอิสระตามสมการ

$$\sigma^2_{(\beta_i)} = \frac{\sigma^2_{(y)}}{s_{ii}(1 - R_i^2)}$$

where

$$s_{ii} = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$$

เราจะเปลี่ยนการหา Regression model ใหม่ โดยแยกค่า Y ออกไป แล้วเปลี่ยน X หนึ่งตัวให้เป็น Y แทนชั่วคราว แล้วทำการวิเคราะห์หา Regression model ระหว่าง X ที่เปลี่ยนมามีฐานะเป็น Y ชั่วคราว กับ X อื่นๆที่เหลือ แล้วนำค่า R^2 (un-adjusted) ที่ได้มาคำนวณหาค่า Variance ของค่าสัมประสิทธ์ แล้วก็เปลี่ยน X ตัวอื่นๆมาเป็น Y ชั่วคราวแทนบ้าง หาค่า Un-adjusted R^2 ของแต่ละ X และคำนวณหาค่า Variance ของค่าสัมประสิทธ์ จนครบทุก X

ถ้าสมมติว่าไม่มีความสัมพันธ์กันเลยระหว่าง X ที่ถูกเปลี่ยนมาเป็น Y ชั่วคราว กับ X ที่เหลืออื่นๆ ค่า Un-adjusted R^2 จะเท่ากับ 0 นั่นคือจะเหลือ

$$\sigma^2_{(\beta_i)} = \frac{\sigma^2_{(y)}}{s_{ii}}$$

แสดงว่าค่า Variance ของค่าสัมประสิทธ์ตัวนั้นๆจะเพิ่มมากขึ้น (เฟ้อ) กว่าที่เป็นอยู่นี้หรือไม่ ขึ้นอยู่กับระดับความสัมพันธ์ของ X ตัวนั้น(ที่เปลี่ยนมามีฐานะเป็น Y ชั่วคราว) กับ X อื่นๆที่เหลือ จะมากน้อยเพียงใด เราเลยเรียกเทอม

นี้ว่า ตัวชี้วัดความเฟ้อ ของ Variance ของค่าสัมประสิทธ์ หรือ Variance Inflation Factor (VIF) มีสมการดังนี้

$$VIF_{(\hat{\beta}_i)} = \frac{1}{1 - R_i^2}$$

จากตารางที่ 1 เมื่อเราให้ X1 มีฐานะเป็นตัวแปรตาม และ X2,X3 และ X4 เป็นตัวแปรอิสระ เมื่อวิเคราะห์ด้วยวิธี Multiple linear regression จะได้ดังต่อไปนี้

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.908
R Square	0.824
Adjusted R Square	0.791
Standard Error	6.050
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	3	2738.591	912.864	24.941	0.000
Residual	16	585.609	36.601		
Total	19	3324.200			

	Coefficients	Standard Error	t Stat	P-value
Intercept	2.376	4.496	0.529	0.604
X2	-0.017	0.188	-0.091	0.928
X3	0.685	0.192	3.559	0.003
X4	0.163	0.273	0.599	0.558

ตารางที่ 4 ผลการวิเคราะห์ Multiple regression โดยโปรแกรม Excel เมื่อ X1 เป็นตัวแปรตาม

จากตารางที่ 4 จะได้

$$VIF_{(\hat{\beta}_1)} = \frac{1}{1 - 0.824} = 5.68$$

เมื่อดูค่า F-Significance และค่า P-Value ของ X2 และ X4 จะพบว่า Model ที่เกิดขึ้นใหม่นี้ยังเกิด Multicollinearity อยู่และ X3 กลายเป็นตัวแปรอิสระที่มีค่านัยสำคัญ ทั้งๆที่ครั้งแรกไม่เป็นเช่นนี้

หากทำการวิเคราะห์ Multiple linear regression เมื่อเปลี่ยน X2,X3และ X4 ไปเป็นตัวแปรตาม แล้วคำนวณหาค่า Variance Inflation Factor จะได้ค่าดังตารางต่อไปนี้

ตัวแปรอิสระ	VIF
X1	5.68
X2	2.50
X3	8.25
X4	4.55

ตารางที่ 5 ค่า VIF ของแต่ละตัวแปรอิสระ

VIF เท่าไหร่ ถึงจะถือว่า Multicollinearity ใน Model นั้นจะเกิดปัญหา

เป็นเรื่องจริงที่ว่า ไม่มีการระบุว่า VIF เท่าใด Multicollinearity จะสร้างปัญหาให้กับการนำ Regression model ที่ได้เมื่อนำไปใช้พยากรณ์ค่าตัวแปรตาม แม้แต่จะสรุปว่าเมื่อเกิด Multicollinearity แล้ว จะแก้ปัญหาอย่างไร จะเกิดความผิดพลาดอะไรบ้าง จะยังสามารถใช้ Model นั้นได้อยู่หรือไม่ ก็ไม่มีตำราที่ไหนเขียนหรือระบุเจาะจงไว้ คงต้องปล่อยให้ผู้ทำการวิเคราะห์ข้อมูลใช้วิจารณญาณส่วนตัวในการจะแก้ปัญหาหรือดำเนินการอย่างหนึ่งอย่างใดต่อไป ถึงแม้ว่าบางตำราจะบอกว่า VIF ตั้งแต่ 10 ขึ้นไป ถือว่า Multicollinearity อาจจะสร้างปัญหาต่อ Regression model ที่ได้ แต่อย่างไรก็ตาม แม้ต่ำกว่า 10 ก็ยังถือว่า สร้างปัญหาได้เช่นกัน

จะทำอย่างไรเมื่อต้องเผชิญกับปัญหา Collinearity หรือ Multicollinearity

วิธีที่ 1 ตัดตัวแปรที่มี Collinearity หรือ Multicollinearity ออกจากการวิเคราะห์หา Regression model จากตัวอย่างที่ผ่านมา พบว่า X3 เป็นตัวแปรที่สมควรตัดออกมากที่สุด ด้วยเหตุผลคือ

- ค่า VIF สูงที่สุด (8.25)

- ค่าสหสัมพันธ์ของ X3 กับ X1 สูงกว่า ค่าสหสัมพันธ์ X3 กับ Y

- เมื่อวิเคราะห์ด้วย Multiple linear regression พบว่า ค่า P-Value ของสัมประสิทธิ์ของ X3 มากกว่า α แสดงว่า

X3 เป็นตัวแปรที่ควรตัดออกจาก Regression model

วิธีที่ 2 รวมตัวแปรที่มี Collinearity กันให้เป็นตัวแปรใหม่ที่ยังให้ความสัมพันธ์กับตัวแปรตามอยู่ เช่นตัวอย่างต่อไปนี้

- ส่วนสูงและน้ำหนัก เป็นตัวแปรอิสระที่มี Correlation กันค่อนข้างมาก เราอาจจะเปลี่ยนไปใช้ตัวแปรใหม่คือ ดัชนีมวลกาย แทน ก็จะตัดปัญหา Collinearity ได้

- ความสูงกับความกว้างของสิ่งของที่เรากำลังศึกษา ถ้ามี Correlation กันค่อนข้างมาก เราอาจจะเปลี่ยนไปใช้ตัวแปรปริมาตร แทน

แต่ก็ไม่ใช่ทุกตัวแปรจะรวมกันได้ อย่างเช่น ระยะเวลาที่อยู่ในครรภ์มารดา (สัปดาห์) และ น้ำหนักทารกแรกเกิด(กก.) ถึงแม้จะมี Correlation กันมาก แต่หากตัดตัวแปรใดตัวแปรหนึ่งออก อาจจะทำให้ Regression model ที่ได้ผิดพลาดมากกว่าที่มีสองตัวแปรนี้อยู่ก็เป็นได้

วิธีที่ 3 ใช้วิธีวิเคราะห์ข้อมูลแบบอื่นที่ไม่สนใจ Collinearity หรือ Multicollinearity เลย เช่น Ridge regression แต่ก็เจอกับความยุ่งยากในการวิเคราะห์มากขึ้นไปอีก เพราะใช้คณิตศาสตร์ค่อนข้างมาก

วิธีที่ 4 ยอมรับว่าต้องมี Collinearity หรือ Multicollinearity แน่ๆ เพราะบางครั้งเราก็ไม่มีทางเลือกที่ดีกว่านี้ ในทางปฏิบัติการที่ Regression model มี Collinearity หรือ Multicollinearity แต่ก็ยังสามารถใช้ในการพยากรณ์ตัวแปร

ตามได้อยู่ เพียงแต่ผู้ใช้ต้องตรวจสอบความถูกต้อง เพิ่มการวิเคราะห์ข้อมูลมากขึ้น

[HOME] [CONTENTS]

การทำ Linear regression model fitting โดยใช้โปรแกรม Microsoft Excel

เนื่องจาก Excel จัดเป็นโปรแกรมที่มีคนใช้งานเยอะและง่ายต่อการใช้งานและ ผู้เขียนคิดว่าผู้อ่านคงจะเคยใช้มาก่อน และมีความชำนาญพอประมาณ ผู้เขียนเลยขอใช้ Excel ในการหา Regression model หรือที่เรียกว่า Model fitting โดยปกติเมื่อเก็บข้อมูลมาแล้ว ก็คงต้องใช้โปรแกรมคอมพิวเตอร์ช่วยในการวิเคราะห์หรือพล็อตกราฟให้ ในการหา Model ที่ดีที่สุดก็เช่นเดียวกัน มีวิธีทำตามขั้นตอนต่อไปนี้

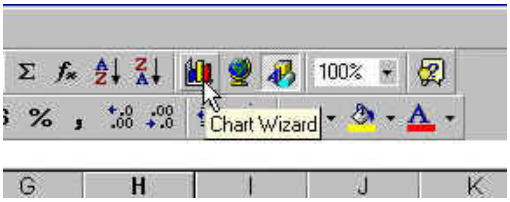
จากข้อมูลของตัวอย่างในเรื่อง Regression Analysis ดังตาราง

ขั้นตอนที่ 1 เมื่อเตรียมข้อมูลในตารางแล้ว เริ่มด้วยการใช้ Mouse เลือกข้อมูลที่จะไปเป็น X,Y ดังรูปที่ 1

	A	B	C	D	E	F	G	H	I	J	K
1	หมีตัวที่	1	2	3	4	5	6	7	8	9	10
2	รอบอก (ซม.)	112	130	148	151	162	189	218	247	315	350
3	น้ำหนัก (กก.)	225	200	459	445	439	577	722	903	1350	1360
4											

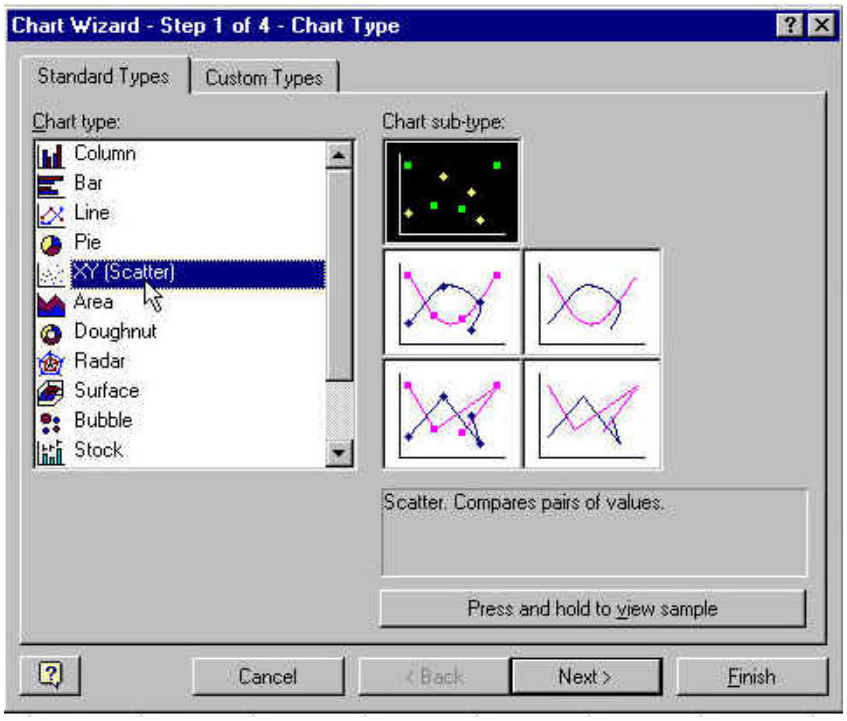
รูปที่ 1

ขั้นตอนที่ 2 ใช้ Mouse กดตรง Chart Wizard ดังรูปที่ 2



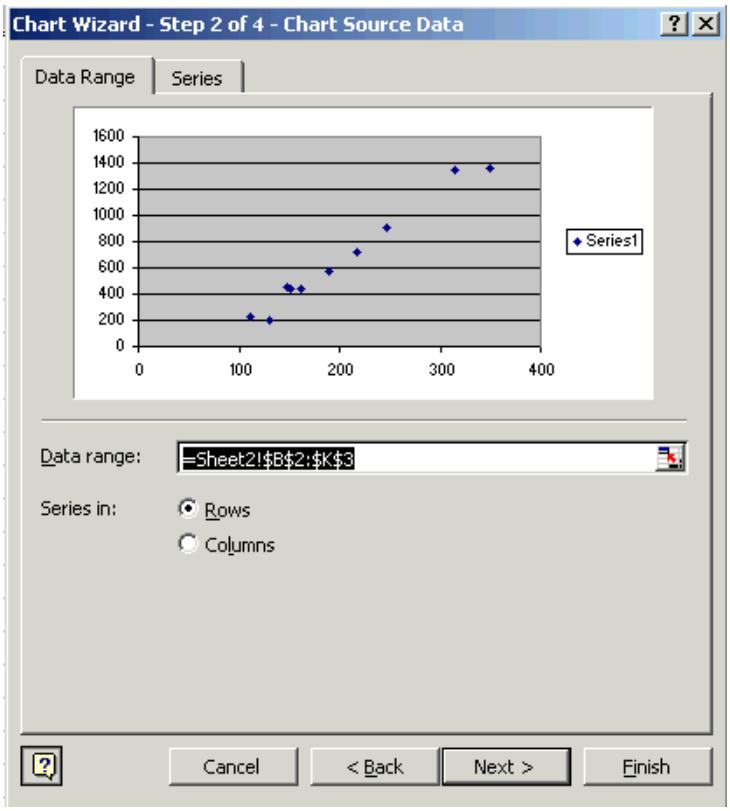
รูปที่ 2

ขั้นตอนที่ 3 เมื่อปรากฏหน้าต่างใหม่ขึ้นมาดังรูปที่ 3 ให้เลือก XY(Scatter) แล้วกด Next



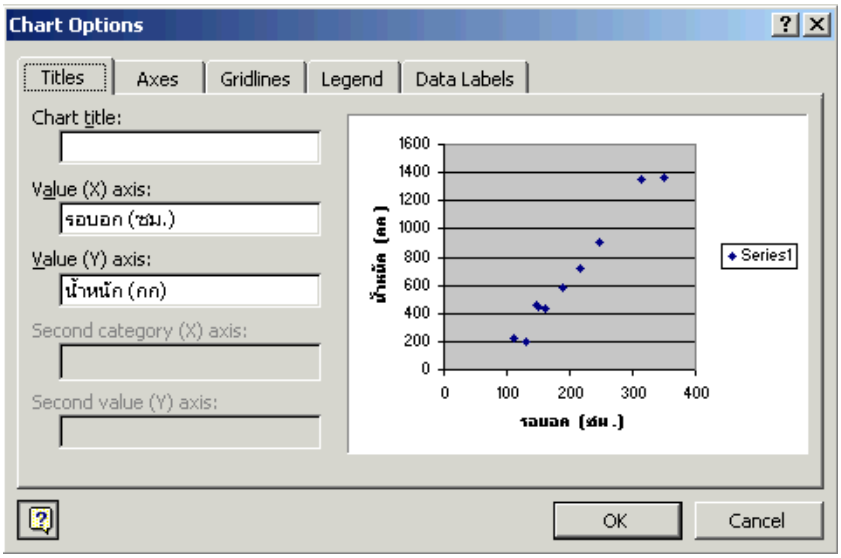
รูปที่ 3

ขั้นตอนที่ 4 เมื่อปรากฏหน้าต่างใหม่ขึ้นมาดังรูปที่ 4 ให้กด Next

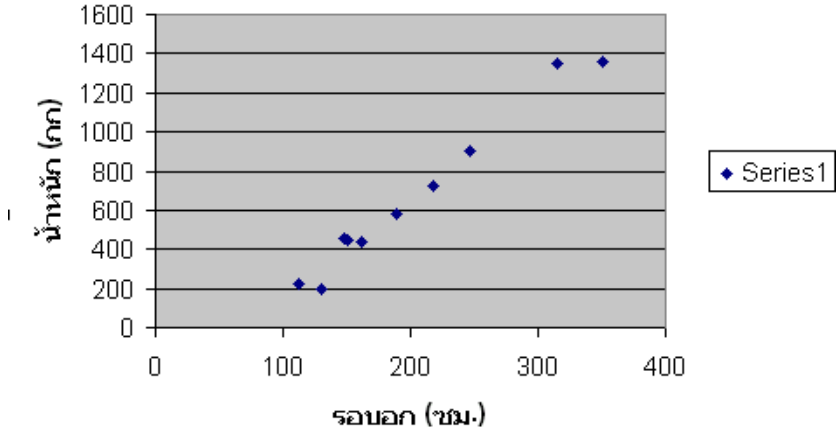


รูปที่ 4

ขั้นตอนที่ 5 เมื่อปรากฏหน้าต่างใหม่ขึ้นมาดังรูปที่ 5 ให้พิมพ์ข้อความที่จะปรากฏตามต้องการแล้วกด OK จะได้กราฟดังรูปที่ 6

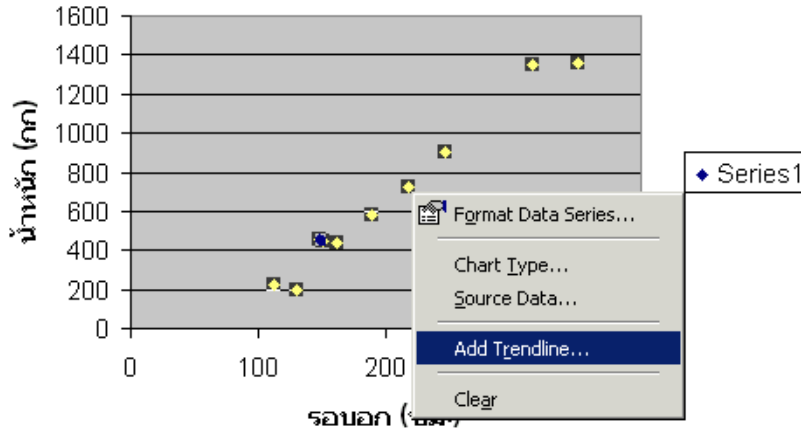


รูปที่ 5

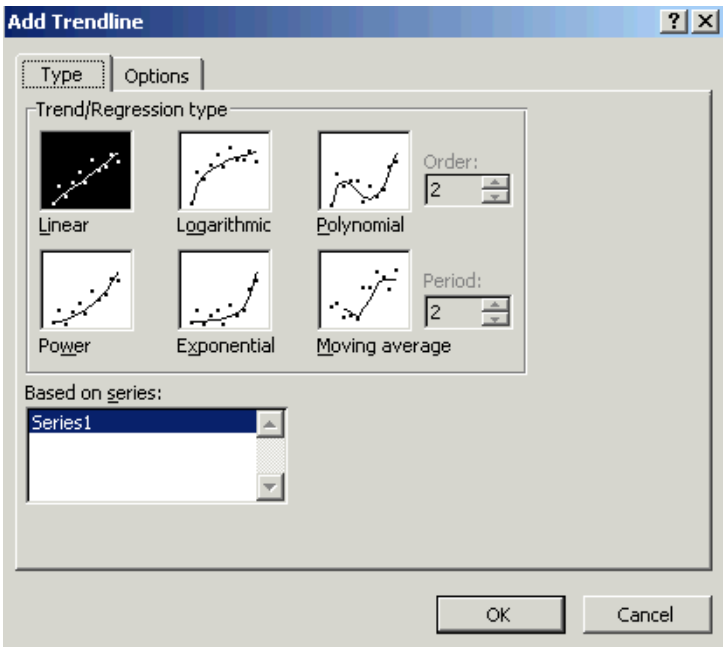


รูปที่ 6

ขั้นตอนที่ 6 เลื่อน mouse cursor ไปแตะจุดที่เรียงตัวกันอยู่ จุดใดจุดหนึ่งก็ได้ จากนั้นให้ Click ขวาที่ Mouse จุดจะเปลี่ยนเป็นสีเหลือง พร้อมกับปรากฏเมนูย่อยๆ ขึ้นมาให้เลือก Add Trendline ก็จะปรากฏหน้าต่างใหม่ดังรูปที่ 8 เช็คดูว่าตรงกราฟ Linear มีสีดำ (แปลว่าเลือก) จากนั้นให้กด OK

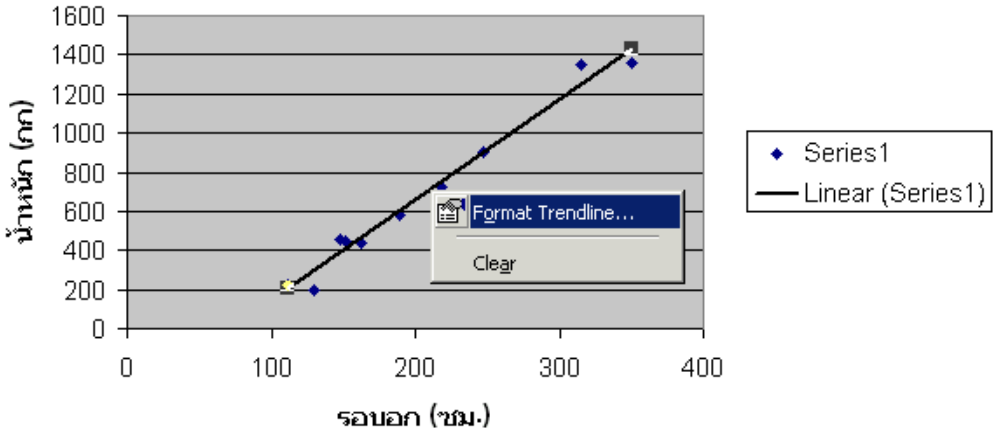


รูปที่ 7

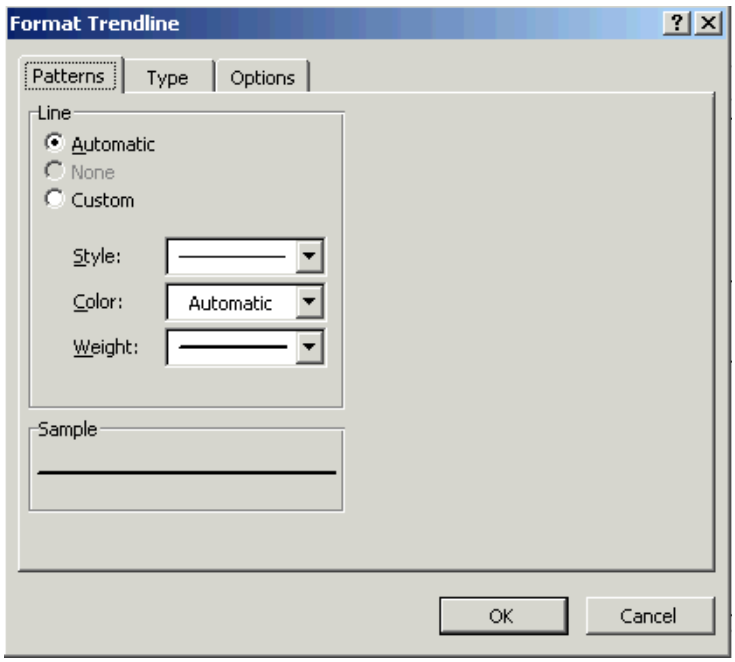


รูปที่ 8

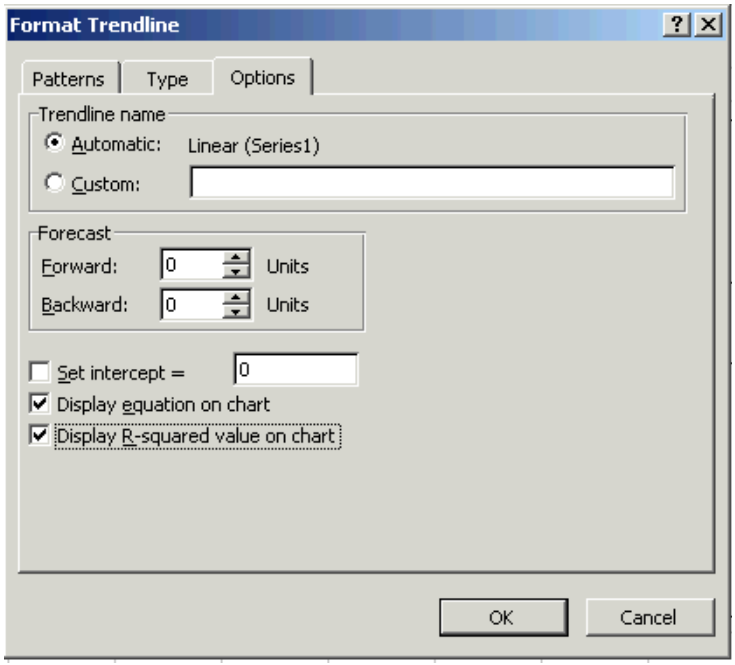
ขั้นตอนที่ 7 เลื่อน mouse cursor ไปแตะกับเส้นตรงที่ปรากฏขึ้นเมื่อเสร็จขั้นตอนที่ 6 จากนั้นให้ Click ขวาที่ Mouse จะปรากฏจุดสีเหลืองที่ปลายเส้นตรงทั้งสองด้าน พร้อมกับปรากฏเมนูย่อยๆ ขึ้นมาให้เลือก Format Trendline ก็จะปรากฏหน้าต่างใหม่ดังรูปที่ 10 ให้กดเลือก Option จะเห็นหน้าต่างเป็นดังรูปที่ 11



รูปที่ 9

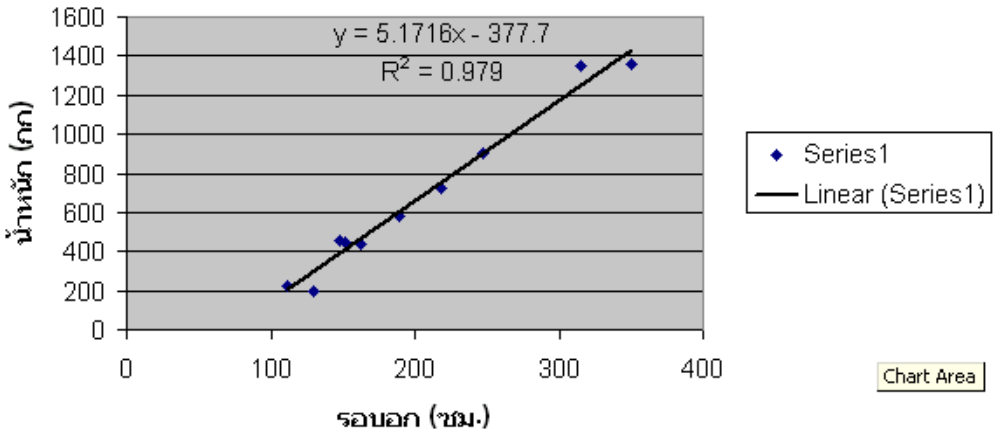


รูปที่ 10



รูปที่ 11

ขั้นตอนที่ 8 ที่หน้าต่างใหม่ตามรูปที่ 11 ให้กดเลือก Display equation on chart และ Display R-Squared value on chart. แล้วกด OK ก็จะปรากฏกรูปกราฟ ดังรูปที่ 12 ที่ปรากฏเพิ่มมาคือ Regression model นั่นคือเราได้ค่า b_0, b_1 และ R-square ซึ่งก็ตรงตามที่คำนวณได้จากหัวข้อที่ผ่านมาทุกประการ



รูปที่ 12

[HOME] [CONTENTS]

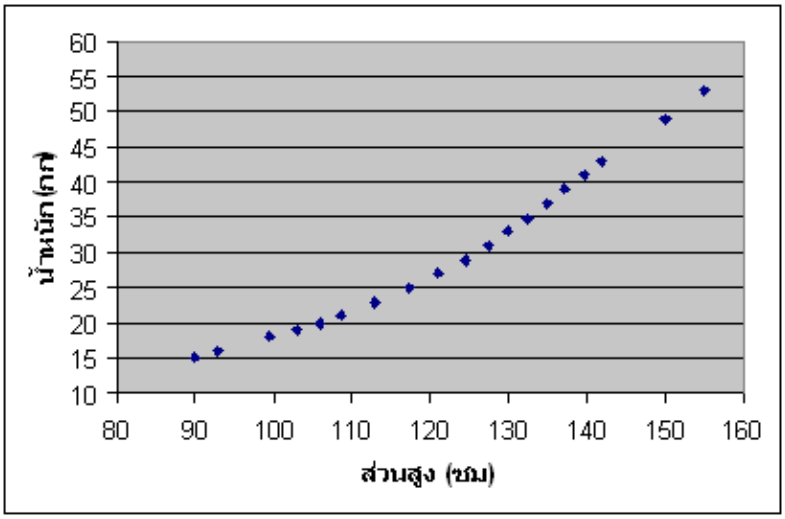
การทำ Polynomial regression model fitting โดยใช้โปรแกรม Microsoft Excel

ต่อเนื่องจากการทำ Linear regression model fitting ซึ่งโดยปกติ ก็มักจะพบอยู่เสมอที่ความสัมพันธ์ระหว่างตัวแปร X,Y ไม่เป็นเชิงเส้น เมื่อทำ Scatter plot ดูแล้วจะเห็นจุดเรียงเป็นแนวโค้ง ซึ่งก็มีหลายระดับหรือหลายดีกรีความโค้งอีกด้วย

ตัวอย่าง 1 ต้องการหา Regression model ความสัมพันธ์ของค่าน้ำหนักและส่วนสูงของ เด็กชายไทย อายุระหว่าง 1-9 ขวบ โดยมีข้อมูลที่ได้จากการวิจัยและเก็บรวบรวมค่าไว้ ดังตัวอย่างในตารางนี้

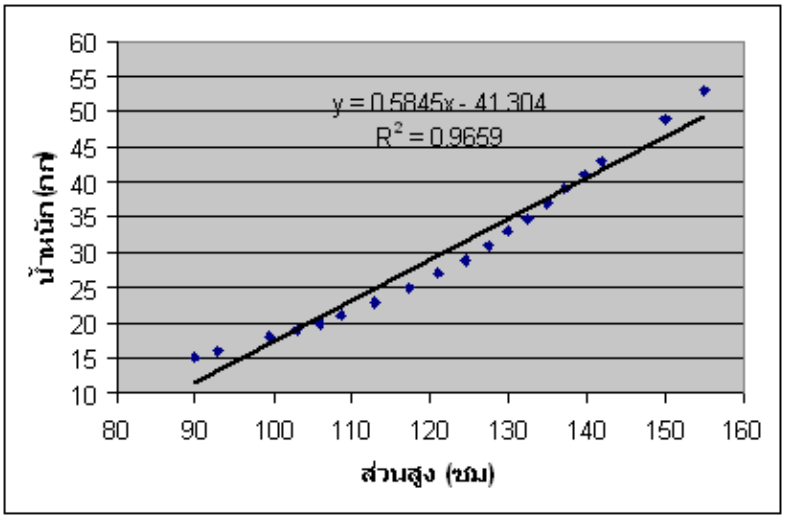
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	ส่วนสูง (ซม)	90	93	100	103	106	109	113	117	121	125	128	130	133	135	137	140	142	150	155	
2	น้ำหนก (กก)	15	16	18	19	20	21	23	25	27	29	31	33	35	37	39	41	43	49	53	
3																					

ขั้นตอนการทำก็จะเหมือนกับกรณี Linear regression ทุกประการ เริ่มจากการทำ Scatter plot ซึ่งข้อมูลจากตาราง เมื่อทำจะได้กราฟดังรูปที่ 1 จะเห็นว่าแนวของจุดจะเป็นเส้นโค้ง



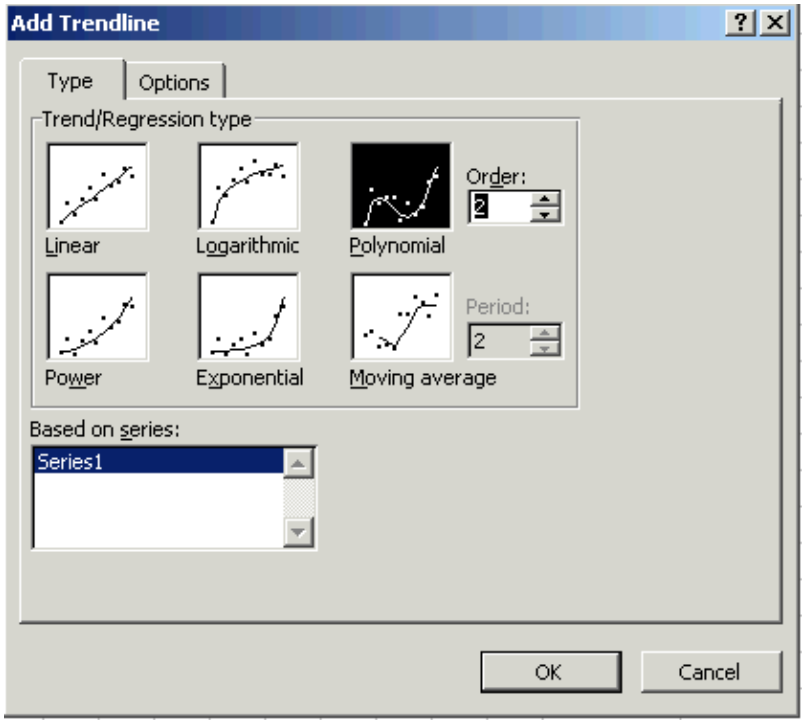
รูปที่ 1

เมื่อถึงขั้นตอน Add trendline ถ้าเราเลือกชนิดเป็น Linear เพื่อลองหา Model ดู กด OK เมื่อเลือกเสร็จแล้ว

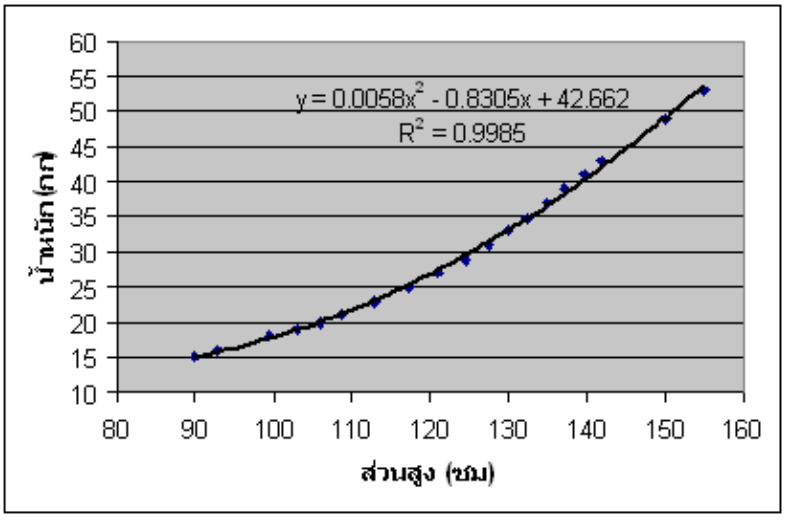


รูปที่ 2 ผลลัพธ์เมื่อเลือก Linear

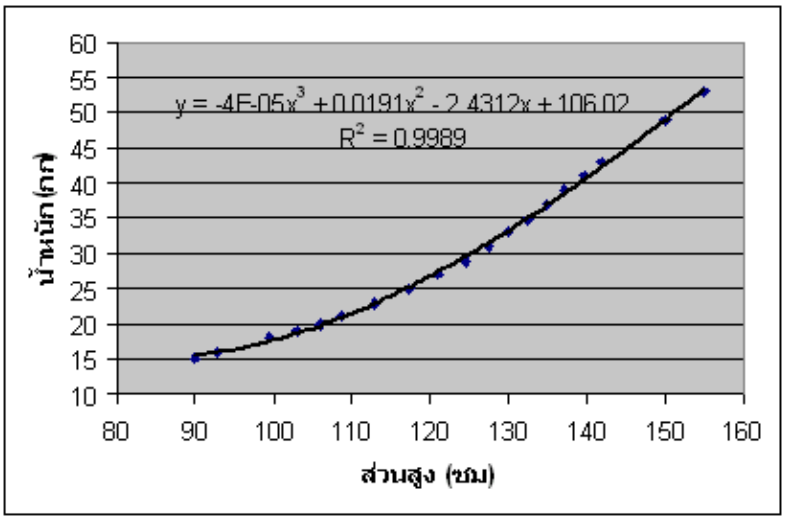
เมื่อถึงขั้นตอน Add trendline หากเลือกชนิดเป็น Polynomial ดังรูปที่ 3 เราจะทดลองเลือก Order เป็น 2 เพื่อลองหา Model ดู กด OK เมื่อเลือกเสร็จแล้ว



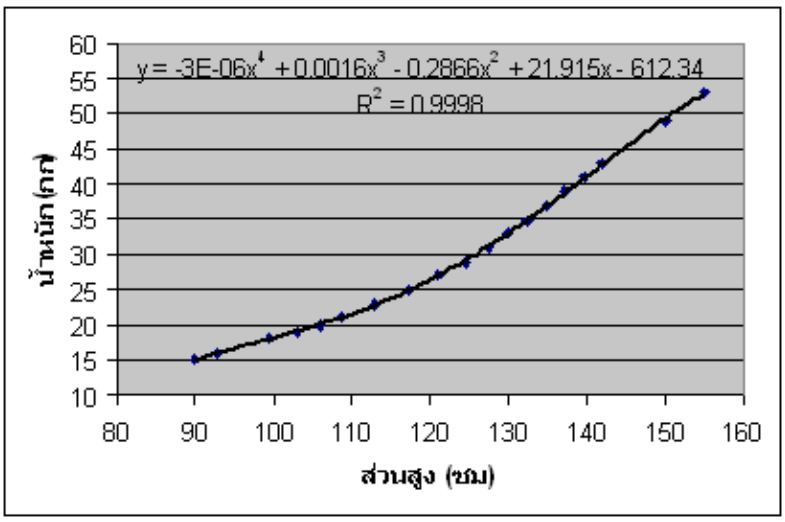
รูปที่ 3



รูปที่ 4 เมื่อเลือก Order 2



รูปที่ 5 เมื่อเลือก Order 3



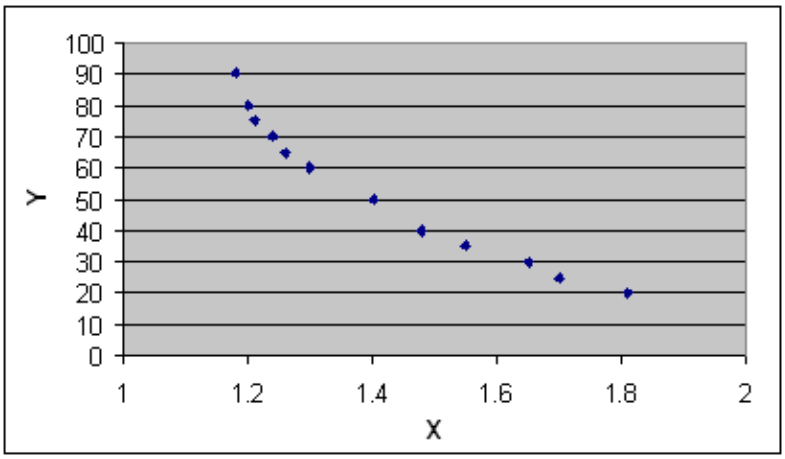
รูปที่ 6 เมื่อเลือก Order 4

จะเห็นว่า เมื่อเราเลือก Model fitting โดยวิธี Linear เส้นตรงจะไม่ทาบทับแนวของจุดได้ดีนัก เมื่อเราดูค่า R-Square ก็จะต่ำกว่า เมื่อเราเลือกใช้วิธี Polynomial ในทำนองเดียวกัน เมื่อเลือกใช้ Polynomial แต่ต่างดีกรี (Order) กัน แนวของเส้นก็จะทาบทับแนวของจุดได้แตกต่างกัน ดูได้จากค่า R-Square แต่ถ้าจะสังเกตดูจะพบว่า แค่ Order 2 ก็เพียงพอแล้วที่จะทำให้ Model ที่ได้มีความเที่ยงตรงสูง เพราะยิ่ง Order สูงขึ้นมากความซับซ้อนของ Model ก็จะยิ่งมากขึ้น แต่ความแม่นยำกลับเพิ่มขึ้นไม่มากนัก จึงไม่มีความจำเป็นที่จะต้องใช้ Order ที่สูงๆ เสมอไป

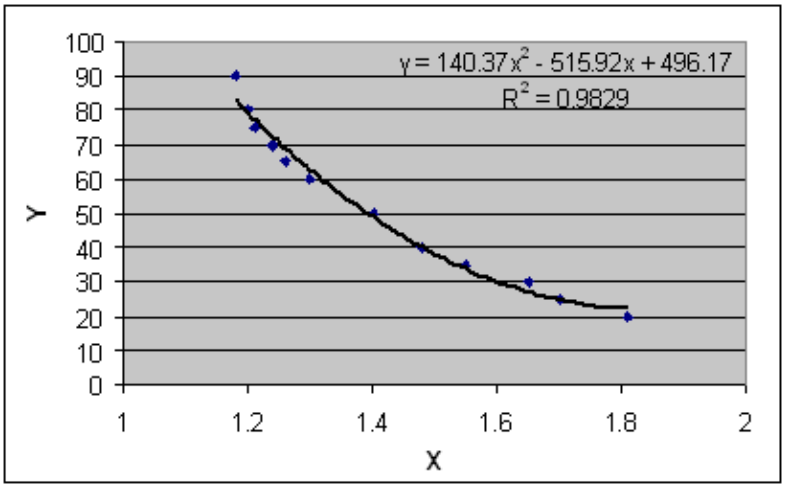
ตัวอย่าง 2 ต้องการหา Regression model ความสัมพันธ์ของ X,Y ในตาราง

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Y	1.81	1.7	1.65	1.55	1.48	1.4	1.3	1.26	1.24	1.21	1.2	1.18
2	X	20	25	30	35	40	50	60	65	70	75	80	90
3													

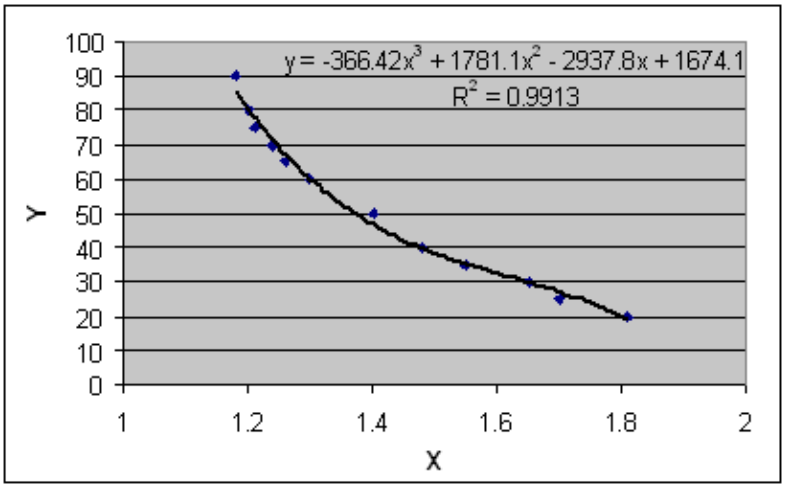
จากการทำ Scatter plot ข้อมูลจากตารางจะได้ กราฟดังรูปที่ 7 จะเห็นว่าแนวของจุดจะเป็นเส้นโค้งลง ไม่เป็นเชิงเส้นแน่นอน



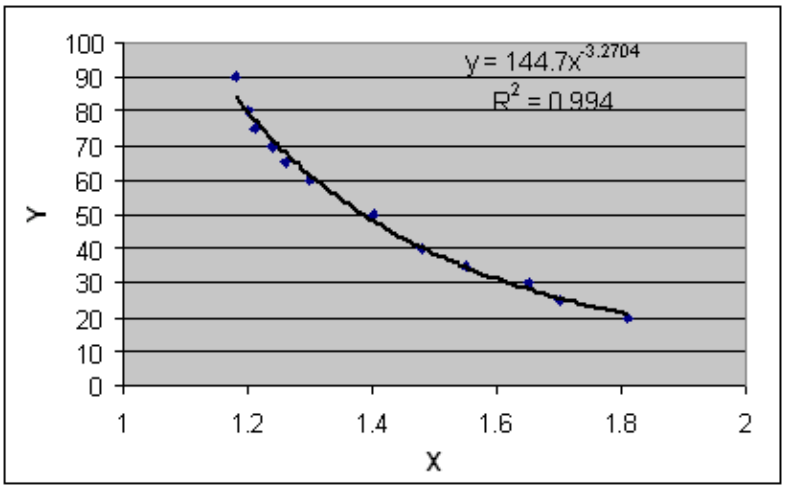
รูปที่ 7



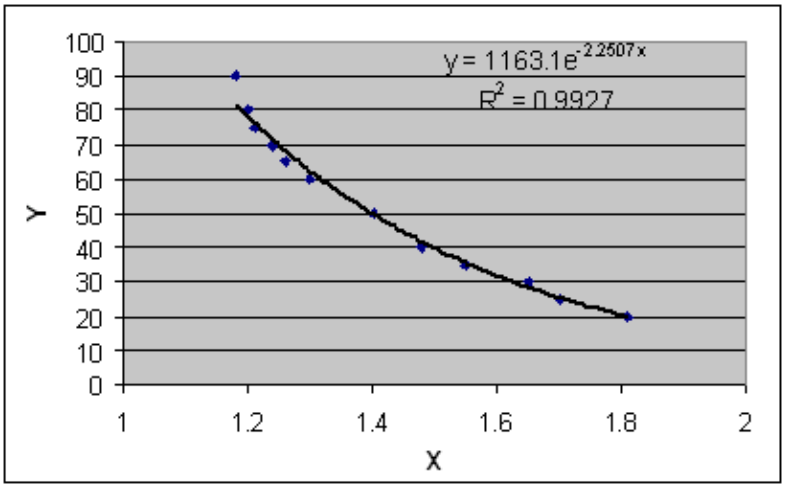
รูปที่ 8 เมื่อเลือกวิธี Polynomial Order 2



รูปที่ 9 เมื่อเลือกวิธี Polynomial Order 3



รูปที่ 10 เมื่อเลือกวิธี Power



รูปที่ 11 เมื่อเลือกวิธี Exponential

ใน Excel มี วิธีการทำ Polynomial regression model fitting อยู่หลายวิธี ในตัวอย่างนี้ผู้เขียนได้เปรียบเทียบให้เห็นความแตกต่างของ Model ที่ได้ ถึงจะมีข้อแตกต่าง แต่เมื่อนำไปใช้พยากรณ์ค่า Y แล้ว ก็จะได้ค่าที่ใกล้เคียงกันมาก สังเกตจากค่า R-Square เมื่อต้องตัดสินใจเลือก Model เราก็ดูความยุ่งยากในการนำไปใช้ เป็นตัวแปรที่สอง ถ้าหากเราคิดว่า R-Square ที่ได้ไม่แตกต่างกันมากนัก

[HOME] [CONTENTS]