การวิเคราะห์การถดถอยพหุคูณ (Multiple regression Analysis)

การวิเคราะห์การถดถอยพหุคูณ (Multiple regression analysis) เป็นเทคนิคที่ใช้ในการหาความสัมพันธ์ระหว่าง ตัวแปรตาม (dependent variable) จำนวนหนึ่งตัว ให้ชื่อว่า Y กับตัวแปรอิสระหรือตัวแปรต้น (independent variables) จำนวนสองตัวหรือมากกว่า ให้ชื่อว่า X(s) ทั้งนี้ก็เพื่อประโยชน์ในการพยากรณ์ ตัวอย่างเช่นในกรณีที่มีตัวแปรต้น 2 ตัว เราอาจแสดงรูปแบบได้ดังนี้

$Y_i = b_O + b_1 X_{1i} + b_2 X_{2i} + u_i$

โดยที่ Y_i = ตัวแปรตาม

b_O = ค่าคงที่ (y-intercept)

b_j = สัมประสิทธิ์หน้าตัวแปรต้น j (j =1,2)มีค่าคงที่

X_{ji} = ตัวแปรต้น j (j=1,2)

u_i = error term

ในกรณีที่มีตัวแปรต้น n ตัวเราอาจแสดงรูปแบบได้ดังนี้

$Y_i = b_O + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + \ldots + b_n X_{ni} + u_i$

และในกรณีที่มีตัวแปรต้นเพียง 1 ตัว เรามีรูปแบบดังนี้

$Y_i = b_O + b_1 X_i + u_i$

ซึ่งรูปแบบนี้เรียกว่า simple regression หรือรูปแบบการถดถอยอย่างง่าย

การประยุกต์ใช้การถดถอยพหุคูณในวงการต่าง ๆ

วงการแพทย์และสุขภาพ: ผู้วิจัยด้านการแพทย์อาจสนใจศึกษาความสัมพันธ์ระหว่างความดันโลหิต (ตัวแปรตาม) กับตัวแปรอื่น ๆ เช่น อายุ เพศ น้ำหนัก อาหารที่รับประทาน การออกกำลังกายและยาที่รับประทาน (ตัวแปรอิสระ) เพื่อพิจารณาความสำคัญของตัวแปรอิสระเหล่านี้ที่มีผลต่อความดันโลหิต

วงการเศรษฐศาสตร์ : นักวิจัยอาจสนใจจะพยากรณ์ผลิตภัณฑ์มวลรวมในประเทศ (Gross Domestic Product-GDP)ที่เป็นตัวแปรตาม จากจำนวนประชากร อัตราเงินเฟ้อ การลงทุนโดยตรงของต่างชาติที่เป็นตัวแปรอิสระ โดยอาศัยเทคนิคการถดถอยพหุคูณ

วงการการศึกษา : นักการศึกษาสามารถศึกษาผลของการเรียนออนไลน์ของนักเรียน/นักศึกษาในช่วงเวลาหนึ่งๆโดยใช้การเปลี่ยนแปลงในผลการเรียนในรูปเกรดที่ได้รับเป็นตัวแปรตาม ส่วนตัวแปรอิสระอาจจะเป็นระดับการขาดความสนใจ พฤติกรรมการติดอินเตอร์เน็ต การไม่ใส่ใจต่อการเรียน

วงการตำรวจจราจร : ผู้วิจัยสามารถศึกษาความสัมพันธ์ระหว่างพฤติกรรมการขับขี่รถที่ไม่ปลอดภัยเช่น การใช้โทรศัพท์ระหว่างขับรถ / การใช้ไลน์ระหว่างขับรถ / การขับรถในยามวิกาล / การเลื่อนเบาะหรือเปลี่ยนรองเท้าในขณะขับรถ ฯลฯเป็นตัวแปรอิสระ กับจำนวนอุบัติเหตุที่เกิดจากผู้ขับขี่ในช่วงเวลาหนึ่งที่เป็นตัวแปรตาม

วงการแพทย์ : นักวิจัยอาจหาความสัมพันธ์ระหว่างน้ำหนัก **(** ตัวแปรตาม **)** กับส่วนสูง อายุ ชั่วโมงที่ออกกำลังกาย **(** ตัวแปรต้น **)** ของบุคคลผู้หนึ่งได้ จากนั้นสามารถใช้ผลลัพธ์ที่ได้จากความสัมพันธ์นี้ไปใช้ในการพยากรณ์ ว่าบุคคลหนึ่งๆมีน้ำหนักมากกว่าค่าเฉลี่ยหรือต่ำกว่าค่าเฉลี่ย เพื่อประโยชน์ในการให้คำปรึกษาแก่ลูกค้าต่อไป

วงการสาธารณสุข : นักระบาดวิทยาที่ทำการศึกษาการระบาดของโรคอาจสนใจศึกษาความสัมพันธ์ระหว่างอัตราการระบาดในกลุ่มประชาชนโดยใช้ข้อมูลจำนวนผู้คนที่เป็นโรคระบาด จากจำนวนประชากรในชุมชนหนึ่งแสนคน (ตัวแปรตาม) และใช้ข้อมูลจากขนาดประชากรในท้องถิ่น ความหนาแน่นของประชากร อุณหภูมิของอากาศ พาหะที่นำพาให้เชื้อโรคแพร่กระจาย ขอบเขตของกลุ่มคนที่มีภูมิคุ้มกันหมู่ (herd immunity) เป็นตัวแปรอิสระ

วงการอสังหาริมทรัพย์ : นักพัฒนาอสังหาริมทรัพย์สามารถศึกษาความสัมพันธ์ระหว่างราคาของบ้าน (ตัวแปรตาม) จากขนาดของบ้าน จำนวนห้องนอน สถานที่ตั้งและปัจจัยอื่น ๆ (ตัวแปรอิสระ) เมื่อทราบความสัมพันธ์แล้ว ก็สามารถใช้กำหนดกลยุทธ์ในการตั้งราคาขายบ้านที่เหมาะสมในโครงการหนึ่งๆ

วงการศึกษา :นักการศึกษาสถาบันอุดมศึกษาสามารถสร้างแบบจำลองเพื่อใช้ในการพยากรณ์คะแนนทดสอบภาษาอังกฤษ (ตัวแปรตาม) โดยพิจารณาจากเพศ อายุ ผลการทดสอบครั้งก่อน ประเภทของโรงเรียน หมวดวิชาที่เลือกเรียนสมัยมัธยม และปัจจัยอื่น ๆ (ตัวแปรต้น)

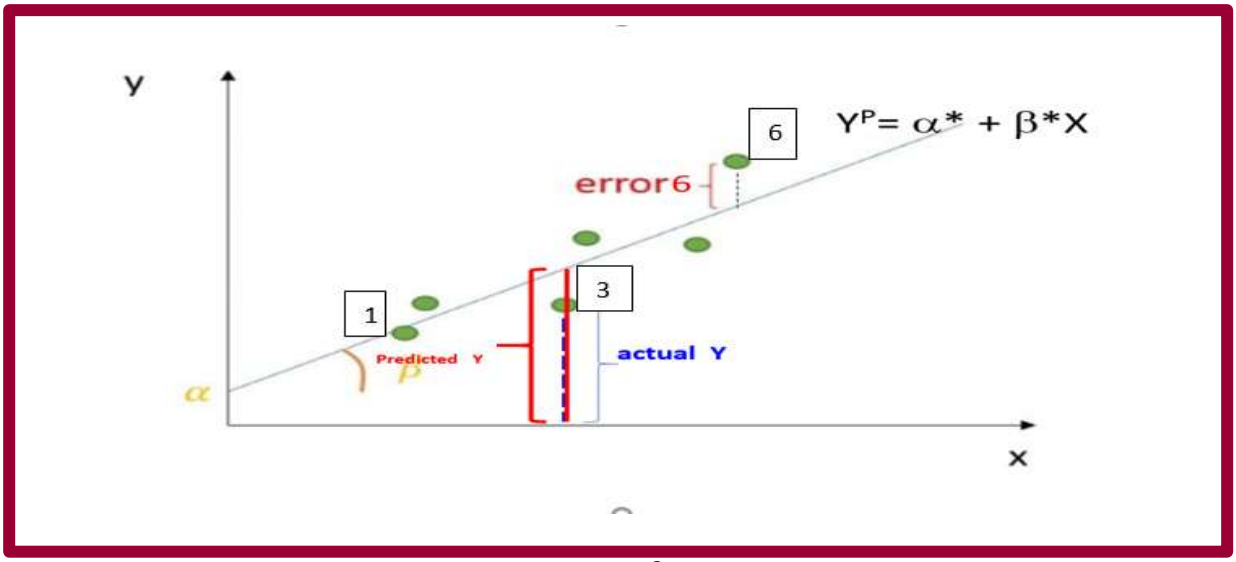
วงการสาธารณสุข: นักวิชาการทางสาธารณสุขสามารถศึกษาอิทธิพลของปัจจัยทางสังคมที่มีผลต่อการเกิดโรคหัวใจในเมือง โดยตัวแปรตามได้แก่ร้อยละของประชากรในเมืองที่มีปัญหาเกี่ยวกับโรคหัวใจ ส่วนตัวแปรอิสระได้แก่ ร้อยละของประชากรในแต่ละเมืองที่สูบบุหรี่ ร้อยละของประชากรในแต่ละเมืองที่ขี่จักรยานไปทำงาน และร้อยละของประชากรในแต่ละเมืองที่เล่นกีฬาเป็นประจำ

วงการการเงินและการลงทุน : นักลงทุนที่ลงทุนในหุ้นที่ออกขายในสาธารณชนและมีความต้องการที่จะกำหนดว่าเมื่อใดที่เป็นช่วงเวลาที่เหมาะสมที่สุดในการขายหุ้น อาจอยากทราบความสัมพันธ์ระหว่างมูลค่าของหุ้น (ตัวแปรตาม) กับปัจจัยหลายอย่างที่มีอิทธิพลต่อการกำหนดราคาหุ้น เช่น ความสามารถในการทำกำไร โครงสร้างต้นทุนการประกอบการ สภาวะการแข่งขัน และโครงสร้างสินทรัพย์ในกิจการ (ตัวแปรอิสระ) ฯลฯ เมื่อสามารถหาความสัมพันธ์ได้แล้วก็สามารถนำไปพยากรณ์ได้ว่าควรจะขายหุ้นออกไปในขณะนี้หรือทำการถือหุ้นนี้ต่อไปอีก

การวิเคราะห์การถดถอยกับ Least Squares Method

เป้าหมายของการวิเคราะห์การถดถอยไม่ว่าจะเป็นในกรณีของการถดถอยอย่างง่าย(simple regression) หรือ การถดถอยพหุคูณ (multiple regression) ก็คือหาสมการของเส้นตรงที่ใช้ตัวแปรต้น (มีหนึ่งตัวในกรณีของการถดถอยอย่างง่าย และหลายตัวในกรณีของการถดถอยพหุคูณ)ในการพยากรณ์ค่าของตัวแปรตาม (Y) ที่มีผลทำให้ผลต่างกำลังสองระหว่างค่า Y ที่แท้จริง (actual Y) กับค่า Y ที่ได้จากการพยากรณ์ (predicted Y หรือ Y^p) มีค่าต่ำที่สุด

ภาพด้านล่างเป็นภาพ 2 มิติแสดงให้เห็นการใช้การวิเคราะห์การถดถอยอย่างง่าย ในการหาสมการเส้นตรงที่มีผลทำให้ $\Sigma (Y- Y^p)^2$ มีค่าต่ำที่สุด



จากภาพสองมิติด้านบนเรามีจุดอยู่ 6 จุดเริ่มจากจุดที่ 1 ทางซ้ายสุด ไปจนถึงจุดที่ 6 ทางขวามือสุด ตามภาพเรามี Y เป็นค่าอยู่บนแกนตั้งและ X เป็นค่าอยู่บนแกนนอน วัตถุประสงค์ก็คือเราต้องการหาเส้นตรงที่ใช้ในการพยากรณ์ค่าของ Y โดยกำหนดค่าของ X ตามภาพเราจะสมมติว่า $Y^p = \alpha^* + \beta^* X$ คือสมการของเส้นตรงเส้นนั้น โปรดสังเกตว่าระยะทางในแนวตั้งลากจากแกน X ขึ้นมาถึงจุดแต่ละจุดคือค่า Y ที่แท้จริง (actual Y) ส่วนระยะทางในแนวตั้งลากจากแกน X ขึ้นมาจนถึงจุดที่เส้น $Y^p=\alpha^*+\beta^* X$ ลากผ่านคือค่า Y ที่ได้จากการพยากรณ์ (predicted Y)

หากเริ่มจากแกนตั้ง Y มองไปทางด้านขวาเราจะเห็นจุด 6 จุด(เลข 1 ใช้กำกับจุดที่หนึ่ง เลข 3 ใช้กำกับจุดที่ 3 และเลข 6 ใช้กำกับจุดที่ 6) จากภาพนี้คงเห็นได้ชัดว่า

ที่จุดที่ 1 ($Y_1 - Y^p_1$) ≈ 0

ที่จุดที่ 2 ($Y_2 - Y^p_2$) มีค่าเป็น +

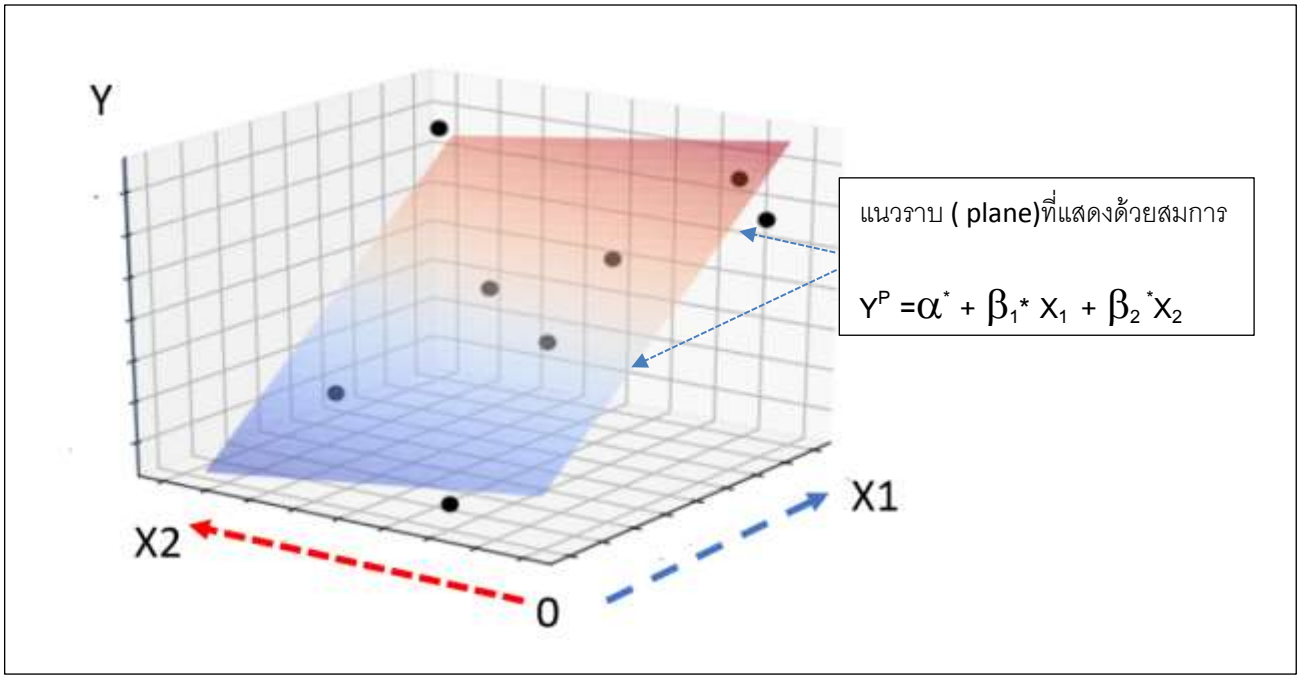
ที่จุดที่ 3 ($Y_3 - Y^p_3$) มีค่าเป็น –

ที่จุดที่ 4 ($Y_4 - Y^p_4$) มีค่าเป็น +

ที่จุดที่ 5 ($Y_5 - Y^P_5$) มีค่าเป็น –

ที่จุดที่ 6 ($Y_6 - Y^p_6$) มีค่าเป็น +

เส้นตรง $Y^p = \alpha^* + \beta^* X$ มีคุณสมบัติที่สำคัญคือเป็นเส้นที่ความผิดพลาด (errors) ในการพยากรณ์มีค่าต่ำที่สุด ไม่มีเส้นใดๆนอกจากนี้จะมีค่าความผิดพลาดต่ำกว่านี้ โดยความผิดพลาดในการพยากรณ์วัดจาก $\sum (Y_i - Y^p{}_i)^2$ จะมีค่าต่ำที่สุด ด้วยเหตุนี้จึงเรียกวิธีการที่ทำให้ได้มาซึ่งสมการเส้นตรงนี้ว่า Least squares method ในกรณีที่เป็นการวิเคราะห์การถดถอยพหุคูณ(multiple regression) ที่มีการใช้ตัวแปรต้นสองตัวในการพยากรณ์ก็เช่นกัน Least squares method สามารถใช้หาสมการของแนวราบ $Y^P = \alpha^* + \beta_1^* X_1 + \beta_2{}^* X_2$ ที่ลากผ่านแนวราบและจุด coordinate (Y, X_1, X_2) แสดงได้ดังรูป 3 มิติต่อไปนี้



ตามภาพแนวราบ (plane) ที่ถูกกำหนดโดยสมการ คือ $Y^* = \alpha^* + \beta^*_1 X_1 + \beta^*_2 X_2$ที่มีผลทำให้ความผิดพลาด (errors) วัดโดยใช้$\Sigma(Y-Y^p)^2$จากจุดทั้ง 8 จุดมีค่าต่ำที่สุด

สมมติฐานที่จำเป็นในการประยุกต์ใช้เทคนิคการถดถอยให้มีประสิทธิภาพ

มีข้อสมมติฐานที่จำเป็นหลายประการที่นักสถิติต้องตระหนักในการใช้เทคนิคการถดถอย

สมมติฐานที่ 1:ตัวแปรตามเป็นตัวแปรที่มีค่าต่อเนื่อง (นั่นก็คือเป็น interval หรือ ratio variable)
สมมติฐานที่ 2:มีตัวแปรอิสระ สองตัวหรือมากกว่าที่เป็นตัวแปรที่มีค่าต่อเนื่อง (เป็น **interval** หรือ **ratio** variable) หรือตัวแปรกำหนดประเภท (เป็น**ordinal** หรือ **nominal** variable)
สมมติฐานที่ 3:ข้อมูลจะต้องไม่มีลักษณะสุดโต่ง (**outliers**) หรือมีการกระจุกตัวแยกออกจากข้อมูลหลัก Tabachnick & Fidell (2001) ได้ให้คำนิยามว่า outliersคือค่าที่มี standardized residual values เกิน 3.3 หรือน้อยกว่า -3.3 หากอยู่นอกขอบเขตนี้อาจมีความจำเป็นที่ต้องตัดทิ้ง หรือแปลงค่าให้อยู่ในรูปคะแนน(scores) โดยต้องระวังไม่ให้ค่าที่ถูกแปลงนั้นมีความแตกต่างเกินไปจากคะแนนอื่น ๆที่เหลือ อนึ่งการกลั่นกรอง Outlier ต้องกระทำทั้งในส่วนที่เป็นตัวแปรต้นและตัวแปรตาม
สมมติฐานที่ 4:ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามมีความสัมพันธ์ในเชิงเส้นตรง ตรวจสอบได้โดยใช้ scatter plot ของค่าตัวแปรอิสระ vs ตัวแปรตาม
สมมติฐานที่ 5:ค่าของ error หรือ residual ต้องมีการกระจายแบบ Normal ตรวจสอบได้โดยดูจาก histogram Normal P-P plot หรือ หรือNormal Q-Q plot ทั้งนี้เพื่อประโยชน์ในการทดสอบสมมติฐานในภายหลัง อนึ่ง หลายท่านอาจเข้าใจผิดว่า ตัวแปรอิสระต้องมีการกระจายแบบ Normal นั้นไม่ถูกต้อง
สมมติฐานที่ 6 : autocorrelation หรือที่มีชื่ออีกชื่อหนึ่งว่า serial correlation คือปัญหาที่นักสถิติจะประสบในการวิเคราะห์การถดถอย (regression analysis) กับข้อมูลซึ่งมีลักษณะเป็นอนุกรมเวลา (time series) โดย error term ในงวดหนึ่งๆมีความสัมพันธ์กับ error term ในงวดก่อน มีผลทำให้ standard errors ที่หาได้มีแนวโน้มต่ำกว่าที่ควรจะเป็น มีผลต่อความน่าเชื่อถือเมื่อใดก็ตามที่ต้องทำการทดสอบข้อสมมติฐานทางสถิติ

ปัญหาการเกิด autocorrelation ไม่ได้จำกัดเฉพาะข้อมูลที่เป็น time-series data เท่านั้น แต่อาจเกิดจากข้อมูลที่เป็นCross-sectional data ที่ซึ่งกลุ่มตัวอย่างอาจมีความสัมพันธ์กันทางใดทางหนึ่ง ตัวอย่างเช่นมีถิ่นฐานทางภูมิศาสตร์เดียวกัน มาจากโรงเรียนเดียวกันเป็นต้น

เราสามารถตรวจสอบปัญหาการมีอยู่ของ autocorrelation ได้โดยดูจาก scatter plot ของ error term กับตัวแปรต้น และจากค่า Durbin-Watson statistic (หมายเหตุ:-โปรดดูรายละเอียดเพิ่มเติมจากStatistics Talks #6)
สมมติฐานที่ 7:ในบรรดาตัวแปรอิสระที่ใช้ในแบบจำลอง จะต้องไม่มีความสัมพันธ์ระหว่างกันที่เรียกว่า Multicollinearity (Pallant,2005 ให้คำนิยามว่า multicollinearity คือตัวแปรสองตัวที่มี coefficient of correlation - r ที่มีค่าตั้งแต่ 0.9 ขึ้นไป) กล่าวคือถ้ามีการใช้ตัวแปรอิสระสองตัว X1 และ X2 ในการพยากรณ์ค่าของตัวแปรตาม Y ค่าของ X1 และ X2 ต้องไม่มีความสัมพันธ์กัน หรือหากมีการใช้ตัวแปรอิสระสามตัว X1 X2 และ X3 ใน

การพยากรณ์ค่าของตัวแปรตาม Y แบบจำลองจะใช้ได้ผลหาก X1 ไม่มีความสัมพันธ์กับ X2 และ X1 ไม่มีความสัมพันธ์กับ X3 และ X2 ก็ไม่มีความสัมพันธ์กับ X3 การตรวจสอบปัญหาการมี multicollinearity เบื้องต้นทำได้โดยดูจาก correlation matrix แต่บางครั้ง correlation matrix ก็ไม่สามารถบ่งบอกว่ามีปัญหา multicollinearity ได้ เพราะอาจเป็นได้ว่าตัวแปรอิสระหลายตัวร่วมกันอาจมีความสัมพันธ์กับตัวแปรตัวอื่นหรือกลุ่มตัวแปรตัวอื่น จำเป็นต้องมีการทดสอบ multicollinearity เป็นการเฉพาะ(Leech, Barett , Morgan , 2005)

ในกรณีที่ตัวแปรอิสระตัวใดตัวหนึ่งเป็น linear combination ของตัวแปรอิสระตัวอื่นอันเป็นผลทำให้ coefficient of correlation-r ระหว่างตัวแปรสองตัวมีค่าเท่ากับ 1 จะเกิดปัญหาที่เรียกว่า singularity ซึ่งจะเป็นผลทำให้เราไม่สามารถหา solutionของปัญหา multiple regression ได้ ตัวอย่างเช่นมีตัวแปรสองตัว ตัวแปรตัวที่หนึ่งบ่งบอกความสูงเป็นนิ้ว ตัวแปรตัวที่สองบ่งบอกความสูงเป็นเซ็นติเมตร จะเห็นได้ว่า ตัวแปรทั้งสองมีความสัมพันธ์กันแบบสมบูรณ์ (perfect relationship) (หมายเหตุ:-โปรดดูรายละเอียดเพิ่มเติมจาก Statistics Talks # 8)

สมมติฐานที่ 8 :ข้อมูลจะต้องมีลักษณะที่เรียกว่า homoscedasticity กล่าวคือ variance ของ error terms จะต้องมีค่าคงที่ตลอดค่าของตัวแปรอิสระ วิธีตรวจสอบก็คือ พิจารณาจาก scatter plot ระหว่าง error terms กับ fitted values (หมายเหตุ:- โปรดดูรายละเอียดเพิ่มเติมจาก Statistics Talks # 9)

สมมติฐานที่ 9: ในการใช้การถดถอยทวีคูณ ต้องมั่นใจว่ากลุ่มตัวอย่างต้องมีขนาดมากเพียงพอ ไม่เล็กจนเกินไปหรือมีข้อมูลที่มีการกระจายที่เบ้ไปทางใดทางหนึ่ง Stevens(1996) กล่าวว่าในการวิจัยด้านสังคมศาสตร์จะต้องมีขนาดกลุ่มตัวอย่าง 15 รายต่อตัวแปรอิสระหนึ่งตัว Tabachnick & Fidell (2001) ได้ให้สูตรในการคำนวณขนาดของกลุ่มตัวอย่างเป็น

$n > 50+8m$ โดย n=ขนาดของกลุ่มตัวอย่าง

m=จำนวนของตัวแปรอิสระ

ดังนั้นหากเรามีตัวแปรอิสระ 5 ตัว ตามเกณฑ์ของ Stevens เราต้องมีขนาดของกลุ่มตัวอย่างเท่ากับ 75 และ 90 ตามเกณฑ์ของ Tabachnick & Fidell

ความหมายของคำศัพท์เฉพาะในเอกสารวิชาการนี้

Coefficient of determination (R^2): สัดส่วนความผันผวนในตัวแปรตาม Y ตามที่อธิบายได้ด้วยแบบจำลอง โดยมีค่าอยู่ระหว่าง 0 ถึง 1 คำนวณได้จากสูตร

$$= \frac{\sum(Y^p - \overline{Y})^2}{\sum(Y - \overline{Y})^2}$$

โดยที่ Y^p = ค่า Y ที่คำนวณได้จากสมการที่ใช้เทคนิคการวิเคราะห์การถดถอย

Y = ค่า Yที่แท้จริง หรือค่า Y ที่ได้จากการสังเกต

$\overline{Y}$ = ค่าเฉลี่ยของตัวแปรตาม

Coefficient of correlation (R): คือค่า square root ของ R^2

Adjusted R^2 : คือค่า coefficient of determination ซึ่งมีการปรับค่าตามจำนวนของตัวแปร โดยเฉพาะอย่างยิ่งจำนวนตัวแปรที่ไม่มีความจำเป็น

P-value: ความน่าจะเป็นที่ค่าสัมประสิทธิ์ตัวใดตัวหนึ่งมีค่าเป็นศูนย์ เมื่อมีการทดสอบความมีนัยสำคัญ

t-**value :** ค่าสถิติที่ใช้ทดสอบค่าสัมประสิทธิ์ตัวใดตัวหนึ่ง

F-statistic: ค่าสถิติที่ใช้ทดสอบความมีนัยสำคัญโดยรวมของแบบจำลอง

Fitted value: ค่า Y ที่ได้จากการพยากรณ์ (Y^p)โดยใช้ตัวแปรอิสระ

Residuals : ผลต่างระหว่างค่า Y ที่ได้จากการสังเกตและค่า Y ที่ได้จากการพยากรณ์ ($Y- Y^p$)

Model-overfitting และ model-underfitting: แบบจำลองที่ over-fitted คือแบบจำลองที่ใช้ได้ดีกับข้อมูลกลุ่มหนึ่ง แต่ไม่สามารถนำไปใช้กับข้อมูลการทดสอบหลายๆข้อมูลกลุ่มใหญ่ทั่วไป โดยหากข้อมูลมีการเปลี่ยนแปลงเพียงเล็กน้อยก็มีผลทำให้ความคลาดเคลื่อนมีสูง

ส่วนแบบจำลองที่ underfitted คือแบบจำลองที่อาจมีการใช้สมมติฐานที่ไม่ถูกต้อง เป็นผลทำให้แบบจำลองนั้นไม่สามารถใช้อธิบายหรือใช้พยากรณ์ความสัมพันธ์ระหว่างข้อมูลได้อย่างมีประสิทธิภาพ ตัวอย่างเช่นในการพยากรณ์น้ำหนักของบุคคลหนึ่งๆ ใช้ความยาวของเท้าวัดจากนิ้วหัวแม่โป้งจนถึงปลายส้นเป็นตัววัด อาจไม่ใช่ข้อมูลที่ดีในการพยากรณ์น้ำหนักของบุคคลนั้น เพราะน้ำหนักอาจขึ้นอยู่กับความสูง ความกว้างของหน้าอก ขนาดของเส้นรอบวงรอบเอว ขนาดของเส้นรอบวงรอบตะโพก

รูปแบบการวิเคราะห์การถดถอยพหุคูณ (Types of multiple regression analysis) : แบ่งออกได้ 3 รูปแบบดังนี้

1.Standard /simultaneous/forced entry multiple regression คือการวิเคราะห์การถดถอยที่ซึ่งตัวแปรอิสระที่ใช้ในแบบจำลองจะถูกนำเข้าสู่แบบจำลองพร้อมกันในคราวเดียว จะมีการประเมิน contribution ในการพยากรณ์ของแต่ละตัวแปรอิสระจากตัวแปรอิสระทั้งหมด เป็นรูปแบบที่ใช้กันแพร่หลายมากที่สุด เป็นวิธีที่ทำให้ทราบว่ากลุ่มตัวแปรอิสระทั้งหมดในรูปกลุ่มมีส่วนช่วยอธิบายความผันผวนในตัวแปรตามมากน้อยเท่าใด

2.Hierarchical หรือ sequential multiple regression คือการวิเคราะห์การถดถอยที่จะมีการพิจารณากำหนดนำตัวแปรอิสระตัวใดตัวหนึ่งหรือกลุ่มใดกลุ่มหนึ่งเข้าสู่แบบจำลองโดยคำนึงถึงหลักการทางทฤษฎีหรือหลักตรรกศาสตร์ ตัวอย่างเช่น ในแวดวงเศรษฐศาสตร์จำนวนผลิตผลทางอุตสาหกรรมจะขึ้นอยู่กับCapitalที่เป็นเครื่องมือเครื่องจักรมากกว่าแรงงาน(Labor) หรือที่ดิน(Land) ผลิตผลทางอุตสาหกรรมเบาที่อาศัยแรงงานเป็นปัจจัยการผลิตหลักจะขึ้นอยู่กับแรงงานเป็นหลัก ในขณะที่ผลิตผลทางการเกษตรจะขึ้นอยู่กับขนาดของที่ดินซึ่งเป็นปัจจัยการผลิตหลักมากกว่าปัจจัยการผลิตอย่างอื่น เมื่อได้มีการนำตัวแปรอิสระหรือกลุ่มตัวแปรอิสระเข้าสู่แบบจำลองแล้ว ก็จะมีการนำตัวแปรอิสระหรือกลุ่มตัวแปรอิสระกลุ่มถัดไปเข้าสู่แบบจำลอง วิธีนี้ผู้วิจัยสามารถประเมิน Contribution ในการพยากรณ์ของตัวแปรอิสระหรือกลุ่มของตัวแปรอิสระหนึ่งๆหลังจากที่ได้มีการควบคุมตัวแปรอิสระหรือกลุ่มตัวแปรอิสระอื่น ๆไว้แล้ว

3.Statistical/stepwise multiple regression คือการวิเคราะห์การถดถอยที่จะมีการกำหนดตัวแปรอิสระที่จะใช้ในแบบจำลอง แต่จะให้Algorithm ของซอฟแวร์ทางสถิติกระทำการคัดเลือกตัวแปรอิสระหนึ่งๆเข้าสู่แบบจำลองโดยอาศัยกฎเกณฑ์ทางสถิติเป็นตัวกำหนด โดยรูปแบบการนำตัวแปรอิสระเข้าหรือออกอาจแบ่งออกเป็น forward selectionที่ซึ่งจะมีการนำตัวแปรอิสระเข้าสู่แบบจำลองทีละตัว Backward deletion ที่ซึ่งจะมีการนำตัวแปรอิสระหลายตัวเข้าสู่แบบจำลองจากนั้นจะมีการคัดตัวแปรอิสระแต่ละตัวออก ส่วนstepwiseจะมีการนำตัวแปรอิสระเข้าสู่แบบจำลองเป็นลำดับขั้น

◉การทดสอบสมมติฐานเกี่ยวกับ regression coefficientตัวใดตัวหนึ่ง

วัตถุประสงค์ :เพื่อใช้ทดสอบว่า Regression coefficient หน้าตัวแปรอิสระตัวใดตัวหนึ่งเป็นศูนย์หรือไม่

ประโยชน์:ทำให้เห็นว่ามีตัวแปรอิสระใดบ้างที่มีส่วนในการอธิบายค่าของตัวแปรตาม

วิธีการ:จะใช้ t-test โดยมีรูปแบบการทดสอบดังนี้ :

$H_O : \beta_i = 0$ (เราสามารถตัดตัวแปรอิสระตัวนี้ออกได้โดยไม่ทำให้ลดคุณค่าในการพยากรณ์)

$H_A : \beta_i \neq 0$ (ตัดตัวแปรนี้ออกไม่ได้ เพราะมีส่วนสำคัญในการพยากรณ์ค่าของตัวแปรตาม)

สูตรที่ใช้ $$t = \frac{b_1 - \beta_i}{S_{bi}}$$

หมายเหตุ:- degree of freedom ที่ใช้คือ n-p-1

โดยที่ n=ขนาดของกลุ่มตัวอย่าง , p=จำนวนตัวแปรอิสระ

S_{bi} = standard error of the regression coefficient

b_1 = ค่าที่ได้จากกลุ่มตัวอย่าง

β_i = ค่าของประชากรหรือตามสมมติฐานที่ตั้งขึ้น

ตัวอย่างเช่น กรณีเรามีตัวแปรอิสระสามตัว X1,X2,X3 เราต้องดำเนินการทดสอบดังต่อไปนี้

$H_O: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$ หากเรายอมรับ H_O แสดงว่า X1 ไม่ช่วยในการอธิบายค่าตัวแปรตาม (Y) ดังนั้นเราอาจตัด X1 ออกจากสมการถดถอยได้	$H_O: \beta_2 = 0 \quad H_A: \beta_2 \neq 0$ หากเรายอมรับ H_O แสดงว่า X2 ไม่ช่วยในการอธิบายค่าตัวแปรตาม (Y) ดังนั้นเราอาจตัด X2 ออกจากสมการถดถอยได้	$H_O: \beta_3 = 0 \quad H_A: \beta_3 \neq 0$ หากเรายอมรับ H_O แสดงว่า X3 ไม่ช่วยในการอธิบายค่าตัวแปรตาม (Y) ดังนั้นเราอาจตัด X3 ออกจากสมการถดถอยได้

◉การทดสอบคุณค่าของสมการถดถอย(regression line) ทั้งเส้น

วัตถุประสงค์ : ใช้ทดสอบว่า regression coefficient หน้าตัวแปรอิสระทุกตัวมีค่าเท่ากับศูนย์หรือไม่

ประโยชน์ : ทำให้เห็นว่าสมการทั้งสมการหรือรูปแบบการวิเคราะห์ที่ดำเนินอยู่มีคุณค่าหรือเอาไปใช้ในการพยากรณ์ค่าของตัวแปรตามได้หรือไม่

วิธีการ: จะใช้ F-testในการทดสอบโดยมีรูปแบบดังนี้

$H_O: \beta_1 = \beta_2 = \beta_3 = \ldots\ldots = 0$ (สมการที่คำนวณได้ไม่มีคุณค่าในการพยากรณ์)

H_A: อย่างน้อยมีค่า β บางค่าที่ไม่ใช่ศูนย์ (สมการที่คำนวณได้มีคุณค่าในการพยากรณ์)

สูตรที่ใช้ $$F = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

โดยที่ p= จำนวนตัวแปรอิสระ ,n =ขนาดของกลุ่มตัวอย่าง

R^2 = Coefficient of determination

หากเรายอมรับ H_O หมายความว่าสมการถดถอยที่ได้ไม่มีคุณค่าในการใช้พยากรณ์ แต่หากเราปฏิเสธ H_O แสดงว่าสมการถดถอยที่ใช้มีคุณค่าในการพยากรณ์

Contribution this issue: ดร.ดนัย ปัตตพงศ์*

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

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

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

* ผู้แต่งหนังสือ"การวิเคราะห์ความเป็นไปได้ทางการเงินและการจัดวงเงินเครดิตของโครงการลงทุน"ประกอบด้วยตัวอย่างของธุรกิจจริงที่ไม่เปิดเผยชื่อนับ 100 บริษัท ครอบคลุมอุตสาหกรรม 24 อุตสาหกรรม

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