รายงาน
เรื่อง การวิเคราะห์การถดถอย
ปัจจัยที่มีอิทธิพลต่อราคาบ้านในเมืองนิวยอร์ก

จัดทำโดย

65030284 นางสาวกัญจนพรรณ ไชยปัญญา

65030292 นางสาวปวีณา โชติประเสริฐ

65030674 นางสาวธนัชพร เครือทอง

เสนอ

ผู้ช่วยศาสตราจารย์ ดร. วนิดา พงษ์ศักดิ์ชาติ

รายงานนี้เป็นส่วนหนึ่งของวิชา 32235165 การวิเคราะห์การถดถอย

สาขาวิทยาการข้อมูลและการวิเคราะห์ข้อมูล

คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา

บทที่ 1
บทนำ

1.1. ที่มาและความสำคัญ

“บ้าน” เป็นสถาปัตยกรรมที่สนองความต้องการพื้นฐานในการดำรงชีวิตของมนุษย์ โดยทั่วไปพฤติกรรมการอยู่อาศัยของมนุษย์มีความต้องการพื้นฐานคือ กิน นอน พักผ่อน ดังนั้นรูปแบบของบ้านจึงถูกสร้างขึ้นเพื่อตอบสนองความต้องการพื้นฐานดังกล่าวในเบื้องต้น และการที่มนุษย์จะเลือกซื้อบ้านนั้นควรพิจารณาถึงปัจจัยหลายๆอย่าง เช่น พื้นที่ของบ้าน ที่ตั้งของบ้าน และประเภทของบ้าน เป็นต้น

นิวยอร์กเป็นเมืองที่มีประชากรหนาแน่นที่สุดในสหรัฐอเมริกา ประมาณ 8.8 ล้านคน เป็นศูนย์กลางเศรษฐกิจและการเงินของโลก มีบริษัทใหญ่ ๆ มากมายตั้งอยู่ในนิวยอร์ก มีระบบขนส่งสาธารณะที่มีประสิทธิภาพ เป็นเมืองที่มีความหลากหลายทางวัฒนธรรม เป็นเมืองที่เต็มไปด้วยความฝันและโอกาส ดึงดูดผู้คนจากทั่วทุกมุมโลกให้มาแสวงหาชีวิตที่ดีกว่า แต่ด้วยความหนาแน่นของประชากรที่คาดการณ์ว่าจะเพิ่มขึ้นอีกในอนาคต บวกกับเศรษฐกิจที่เฟื่องฟู ส่งผลให้ “บ้าน” กลายเป็นสิ่งล้ำค่าที่หลายคนใฝ่ฝัน จึงทำให้บ้านมีราคาเพิ่มสูงขึ้นเรื่อย ๆ ซึ่งความต้องการบ้านในนิวยอร์กนั้นมีหลากหลาย ขึ้นอยู่กับไลฟ์สไตล์ งบประมาณ พื้นที่ใช้สอย สถานที่ตั้ง และความต้องการของแต่ละบุคคล

ดังนั้น เราจึงอยากทราบว่ามีปัจจัยใดบ้างที่มีผลต่อราคาบ้านในเมืองนิวยอร์ก ซึ่งพิจารณาจากจำนวนห้องนอน จำนวนห้องน้ำ พื้นที่บ้านรวมที่ดิน ประเภทของบ้าน และเขตที่ตั้งของบ้าน โดยใช้การสร้างตัวแบบที่เหมาะสมสำหรับการพยากรณ์ราคาบ้านในเมืองนิวยอร์ก

1.2. ข้อมูล (Data)

ข้อมูลที่นำมาใช้ในการศึกษาการวิเคราะห์การถดถอยครั้งนี้เป็นข้อมูลประเภททุติยภูมิ ซึ่งเป็นข้อมูลที่ได้มาจากการเก็บรวบรวมอยู่ในเว็บไซต์ Kaggle โดย NIDULA ELGIRIYEWITHANA ซึ่งคณะผู้จัดทำได้นำข้อมูลมาทำ Data Cleansing แล้วได้ข้อมูลจำนวน 4,383 แถว 6 คอลัมน์ มีลักษณะดังนี้

i	PRICE	BEDS	BATH	PROPERTYSQFT	TYPE	COUNTY
1	315,000	2	2	1400.00	Condo	Manhattan
2	19,500,000	7	10	17545.00	Condo	Manhattan
3	69,000	3	1	445.00	Condo	Manhattan
4	899,500	2	2	2184.21	Condo	Manhattan
5	260,000	4	2	2015.00	House	Staten Island
6	690,000	5	2	4004.00	House	Brooklyn
7	55,000,000	7	2	14175.00	Townhouse	Manhattan
8	16,800,000	8	16	33000.00	House	Staten Island
9	265,000	1	1	750.00	Co-op	The Bronx
10	440,000	2	1	978.00	Co-op	Brooklyn

ตารางที่ 1 ตารางตัวอย่างข้อมูลที่ใช้ในการวิเคราะห์การถดถอย

โดยรายละเอียดของข้อมูล มีดังนี้

1. PRICE คือ ราคาบ้านในเมืองนิวยอร์ก (ดอลลาร์สหรัฐ) เป็นข้อมูลเชิงปริมาณ
2. BEDS คือ จำนวนห้องนอน (ห้อง) เป็นข้อมูลเชิงปริมาณ
3. BATH คือ จำนวนห้องน้ำ (ห้อง) เป็นข้อมูลเชิงปริมาณ
4. PROPERTYSQFT คือ พื้นที่บ้านรวมที่ดิน (ตารางฟุต) เป็นข้อมูลเชิงปริมาณ
5. TYPE คือ ประเภทของบ้าน (ประเภท) เป็นข้อมูลเชิงคุณภาพ ซึ่งประกอบไปด้วย Co-op, Condo, House, Multi-family home และ Townhouse
6. COUNTY คือ เขตของบ้านในเมืองนิวยอร์ก เป็นข้อมูลเชิงคุณภาพ ซึ่งประกอบไปด้วย Brooklyn, Manhattan, Queens, Staten Island และ The Bronx

บทที่ 2
ขั้นตอนการศึกษา

ในบทนี้จะกล่าวถึงขั้นตอนที่ใช้ในการวิเคราะห์การถดถอยของปัจจัยที่มีอิทธิพลต่อราคาบ้านในเมืองนิวยอร์ก ได้แก่ จำนวนห้องนอน จำนวนห้องน้ำ พื้นที่บ้านรวมที่ดิน ประเภทของบ้าน และเขตที่ตั้งของบ้าน

2.1 ข้อมูลที่ใช้ในการศึกษา

ข้อมูลที่นำมาใช้ในการศึกษาการวิเคราะห์การถดถอยครั้งนี้เป็นข้อมูลประเภททุติยภูมิ ซึ่งเป็นข้อมูลที่ได้มาจากการเก็บรวบรวมอยู่ในเว็บไซต์ Kaggle โดย NIDULA ELGIRIYEWITHANA ซึ่งข้อมูลมีจำนวน 4,383 แถว 6 คอลัมน์ มีลักษณะดังนี้

i	PRICE	BEDS	BATH	PROPERTYSQFT	TYPE	COUNTY
1	315,000	2	2	1400.00	Condo	Manhattan
2	19,500,000	7	10	17545.00	Condo	Manhattan
3	69,000	3	1	445.00	Condo	Manhattan
4	899,500	2	2	2184.21	Condo	Manhattan
5	260,000	4	2	2015.00	House	Staten Island
6	690,000	5	2	4004.00	House	Brooklyn
7	55,000,000	7	2	14175.00	Townhouse	Manhattan
8	16,800,000	8	16	33000.00	House	Staten Island
9	265,000	1	1	750.00	Co-op	The Bronx
10	440,000	2	1	978.00	Co-op	Brooklyn

ตารางที่ 1 ตารางตัวอย่างข้อมูลที่ใช้ในการวิเคราะห์การถดถอย

โดยรายละเอียดของข้อมูล มีดังนี้

1. PRICE คือ ราคาบ้านในเมืองนิวยอร์ก (ดอลลาร์สหรัฐ) เป็นข้อมูลเชิงปริมาณ
2. BEDS คือ จำนวนห้องนอน (ห้อง) เป็นข้อมูลเชิงปริมาณ
3. BATH คือ จำนวนห้องน้ำ (ห้อง) เป็นข้อมูลเชิงปริมาณ
4. PROPERTYSQFT คือ พื้นที่บ้านรวมที่ดิน (ตารางฟุต) เป็นข้อมูลเชิงปริมาณ
5. TYPE คือ ประเภทของบ้าน (ประเภท) เป็นข้อมูลเชิงคุณภาพ ซึ่งประกอบไปด้วย
   Co-op, Condo, House, Multi-family home และ Townhouse

6. COUNTY คือ เขตของบ้านในเมืองนิวยอร์ก เป็นข้อมูลเชิงคุณภาพ ซึ่งประกอบไปด้วย Brooklyn, Manhattan, Queens, Staten Island และ The Bronx

2.2 ศึกษาข้อมูลเบื้องต้นเกี่ยวกับลักษณะของข้อมูล

การศึกษาข้อมูลเบื้องต้น เป็นการศึกษาจากค่าสถิติพื้นฐาน เช่น ค่าเฉลี่ย ค่าสูงสุด ค่าต่ำสุด ค่าเบี่ยงเบนมาตรฐาน เป็นต้น เพื่อพิจารณาลักษณะของข้อมูลของตัวแปรแต่ละตัวแปรที่สนใจศึกษา โดยการศึกษาข้อมูลเบื้องต้นนี้ได้มีการแบ่งการศึกษาออกเป็น 2 ส่วน คือ ส่วนของตัวแปรเชิงปริมาณ และส่วนของตัวแปรเชิงคุณภาพ ดังนี้

2.2.1 การศึกษาข้อมูลเบื้องต้นของตัวแปรเชิงปริมาณ

Variable	N	Mean	Median	Variance	Std. Dev.	min	max
PRICE	4,383	1,933,627	848,000	17,632,570,000,000	4,199,116	49,500	60,000,000
BEDS	4,383	3.4	3	7.16584	2.7	1	50
BATH	4,383	2.4	2	4.00708	2	0	50
PROPERTYSQST	4,383	2,197	2,184.208	6,040,942	2,458	250	65,535

ตารางที่ 2 ตารางค่าสถิติพื้นฐานของตัวแปรเชิงปริมาณ

การศึกษาค่าสถิติพื้นฐานของตัวแปรเชิงปริมาณเบื้องต้น พบว่า ราคาบ้านในเมืองนิวยอร์ก (PRICE) ที่ศึกษานี้ มีค่าเฉลี่ยอยู่ 1,933,627 ดอลลาร์สหรัฐ และมีการกระจายข้อมูลตั้งแต่ 49,500 ถึง 60,000,000 ดอลลาร์สหรัฐ ดังรูป ก. (ภาคผนวก ข.) ค่าเฉลี่ยของจำนวนห้องนอน (BEDS) คือ 3.4 ห้อง หรือประมาณ 3 ห้องนอน โดยข้อมูลจะมีการกระจายตั้งแต่ 1 ถึง 50 ห้องนอน ดังรูป ข. (ภาคผนวก ข.) ส่วนค่าเฉลี่ยของห้องน้ำ (BATH) นั้นคือ 2.4 ห้อง หรือประมาณ 2 ห้อง ซึ่งข้อมูลจะกระจายตั้งแต่ 0 ถึง 50 ห้อง ดังรูป ค. (ภาคผนวก ข.) และค่าเฉลี่ยของพื้นที่บ้านรวมที่ดิน (PROPERTYSQFT) จะมีค่า 2,184.208 ตารางฟุต มีการกระจายตัวของข้อมูลตั้งแต่ 250 ถึง 65,535 ตารางฟุต ดังรูป ง. (ภาคผนวก ข.)

2.2.2 การศึกษาข้อมูลเบื้องต้นของตัวแปรเชิงคุณภาพ

Variable	N	
TYPE		
... Co-op	1,450	33%
... Condo	896	20%
... House	1,011	23%
... Multi-family home	727	17%
... Townhouse	299	7%
COUNTY		
... Brooklyn	1,101	25%
... Manhattan	1,214	28%
... Queens	1,165	27%
... Staten Island	429	10%
... The Bronx	474	10%

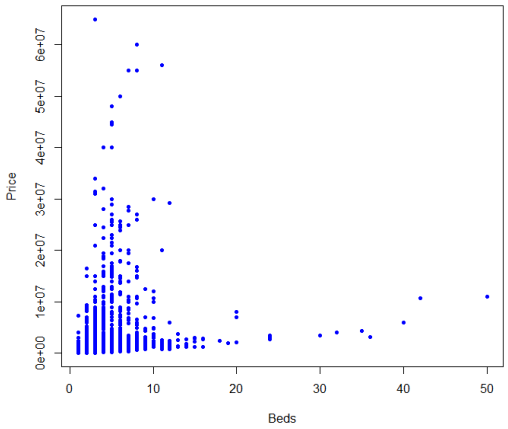
ตารางที่ 3 ตารางสรุปข้อมูลของตัวแปรเชิงคุณภาพ

การศึกษาเบื้องต้นของตัวแปรเชิงคุณภาพ พบว่า ประเภทของบ้าน (TYPE) แบ่งเป็น 5 กลุ่ม คือ ประเภท Co-op จำนวน 1,450 ค่า คิดเป็น 33% ประเภท Condo จำนวน 896 ค่า คิดเป็น 20% ประเภท House จำนวน 1,011 ค่า คิดเป็น 23% ประเภท Multi-family home จำนวน 727 ค่า คิดเป็น 17% และประเภท Townhouse จำนวน 299 ค่า คิดเป็น 7% ดังรูป จ. (ภาคผนวก ข.) ส่วนเขตของบ้าน (COUNTY) แบ่งออกเป็น 5 กลุ่ม คือ Brooklyn จำนวน 1,101 ค่า คิดเป็น 25% Manhattan จำนวน 1,214 ค่า คิดเป็น 28% Queens จำนวน 1,165 ค่า คิดเป็น 27% Staten-Island จำนวน 429 ค่า คิดเป็น 10% และ The Bronx จำนวน 474 ค่า คิดเป็น 10% ดังรูป ฉ. (ภาคผนวก ข.)

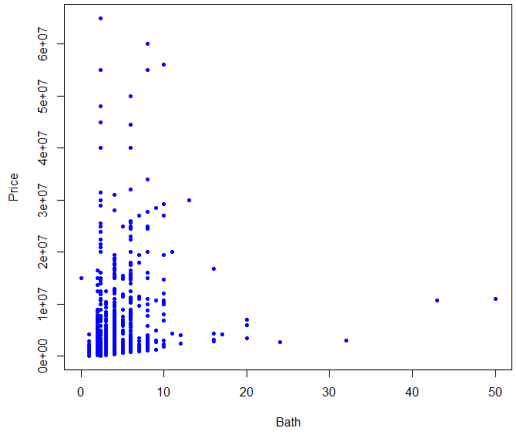
2.3 พิจารณารูปแบบความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระแต่ละตัวแปร โดยใช้แผนภาพการกระจาย

ก่อนจะนำข้อมูลตัวแปรแต่ละตัวมาสร้างกราฟดูความสัมพันธ์เชิงเส้นนั้น ได้มีการกำหนดให้ ราคาบ้านในเมืองนิวยอร์ก (PRICE) เป็นตัวแปรตาม และจำนวนห้องนอน (BEDS) จำนวนห้องน้ำ (BATH) พื้นที่บ้านรวมที่ดิน (PROPERTYSQFT) ประเภทของบ้าน (TYPE) และเขตที่ตั้งของบ้านในเมืองนิวยอร์ก (COUNTY) เป็นตัวแปรอิสระ

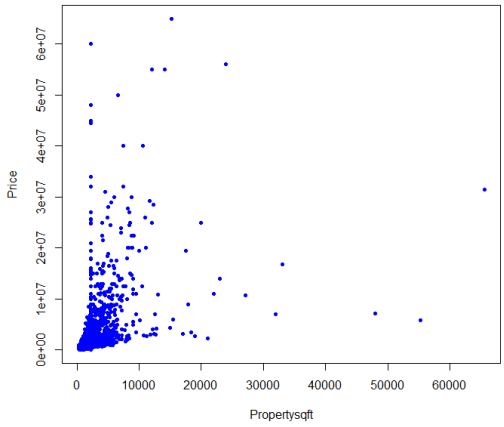
จากนั้นนำตัวแปรอิสระเชิงปริมาณแต่ละตัวมาสร้างแผนภาพเพื่อดูความสัมพันธ์กับตัวแปรตามเป็นคู่ ๆ เพื่อดูแนวโน้มของความสัมพันธ์เชิงเส้น ซึ่งได้แผนภาพความสัมพันธ์ออกมาดังรูปต่อไปนี้



(a) แผนภาพการกระจายระหว่างราคาบ้านและจำนวนห้องนอน



(b) แผนภาพการกระจายระหว่างราคาบ้านและจำนวนห้องน้ำ

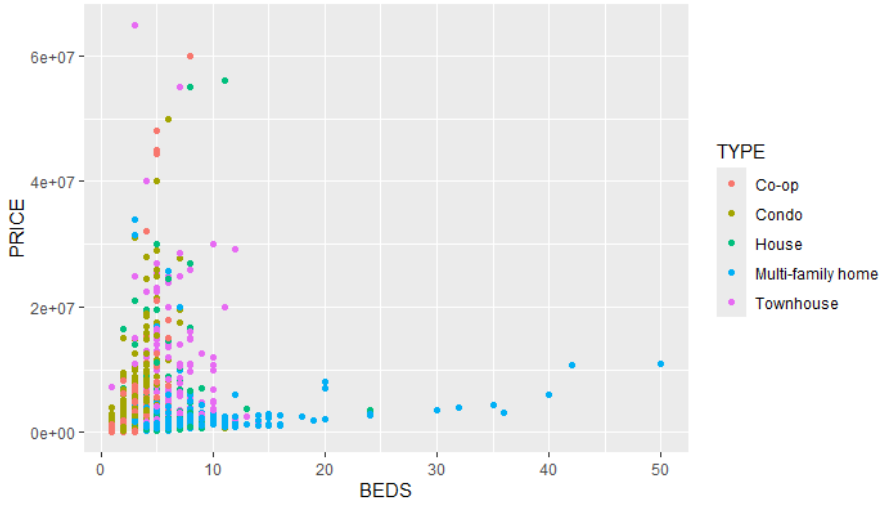


(c) แผนภาพการกระจายระหว่างราคาบ้านและพื้นที่บ้านรวมที่ดิน

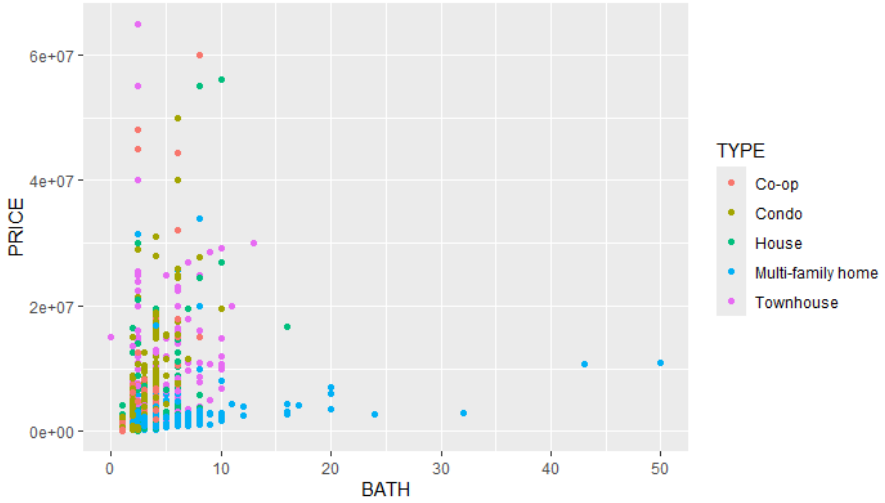
รูปที่ 1 แผนภาพการกระจายระหว่างตัวแปรตามและตัวแปรอิสระเชิงปริมาณ

ดังนั้นจากแผนภาพการกระจายระหว่างตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณกับตัวแปรตามดังรูปที่ 1 พบว่าราคาบ้านในเมืองนิวยอร์กมีความสัมพันธ์เชิงเส้นกับจำนวนห้องนอน ดังรูปที่ 1(a) จำนวนห้องน้ำ ดังรูปที่ 1(b) และพื้นที่บ้านรวมที่ดิน ดังรูปที่ 1(c) โดยทิศทางความสัมพันธ์เป็นไปในทิศทางเดียวกัน

แต่ในที่นี้เราไม่สามารถสร้างแผนภาพแสดงความสัมพันธ์ระหว่างตัวแปรเชิงคุณภาพกับตัวแปรตามได้ จึงต้องอาศัยการสร้างแผนภาพความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระเชิงปริมาณ แล้วใช้ตัวแปรอิสระเชิงคุณภาพแบ่งกลุ่มการกระจายของข้อมูล ซึ่งจะได้แผนภาพความสัมพันธ์ดังต่อไปนี้



(a) แผนภาพการกระจายระหว่างราคาบ้านและจำนวนห้องนอน โดยแบ่งตามประเภทของบ้าน

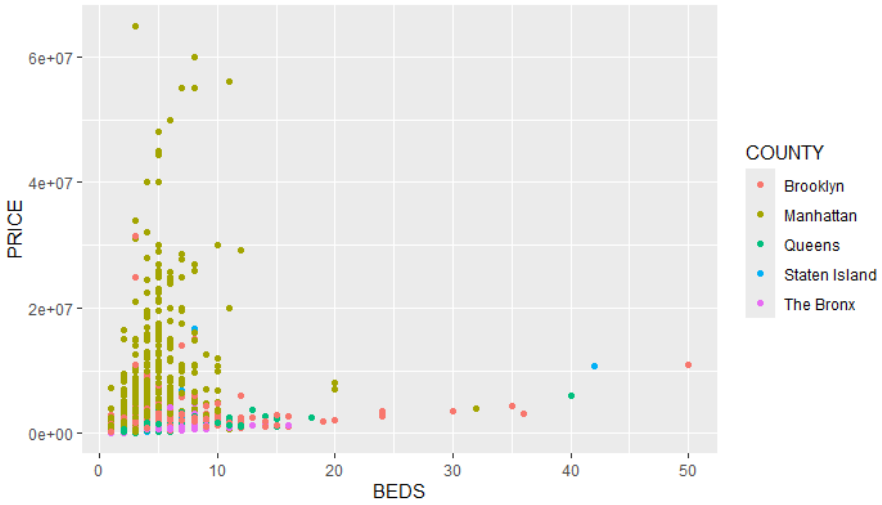


(b) แผนภาพการกระจายระหว่างราคาบ้านและจำนวนห้องน้ำ โดยแบ่งตามประเภทของบ้าน

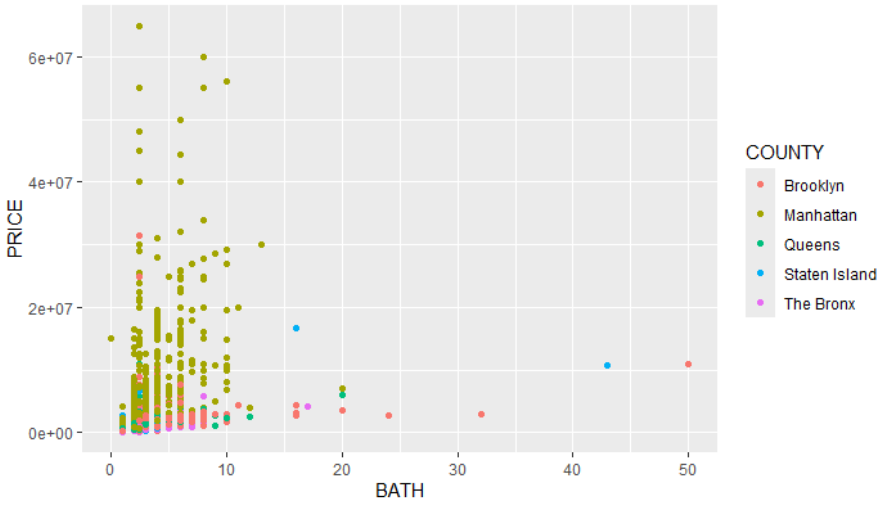
(c) แผนภาพการกระจายระหว่างราคาบ้านและพื้นที่บ้านรวมที่ดิน โดยแบ่งตามประเภทของบ้าน

รูปที่ 2 แผนภาพการกระจายระหว่างราคาบ้านและตัวแปรอิสระเชิงปริมาณ โดยแบ่งตามประเภทของบ้าน

จากแผนภาพการกระจายระหว่างราคาบ้านและตัวแปรอิสระเชิงปริมาณ โดยแบ่งตามประเภทของบ้าน ดังรูปที่ 2 จะเห็นว่าข้อมูลมีการกระจายเป็น 5 กลุ่มตามประเภทของบ้าน โดยบ้านประเภท Multi-family home ข้อมูลมีการกระจายในช่วงราคาที่ต่ำเมื่อเทียบกับบ้านประเภทอื่น ๆ



(a) แผนภาพการกระจายระหว่างราคาบ้านและจำนวนห้องนอน โดยแบ่งตามเขตที่ตั้งของบ้าน



(b) แผนภาพการกระจายระหว่างราคาบ้านและจำนวนห้องน้ำ โดยแบ่งตามเขตที่ตั้งของบ้าน

(c) แผนภาพการกระจายระหว่างราคาบ้านและพื้นที่บ้านรวมที่ดิน โดยแบ่งตามเขตที่ตั้งของบ้าน

รูปที่ 3 แผนภาพการกระจายระหว่างราคาบ้านและตัวแปรอิสระเชิงปริมาณ โดยแบ่งตามเขตที่ตั้งของบ้าน

จากแผนภาพการกระจายระหว่างราคาบ้านและตัวแปรอิสระเชิงปริมาณ โดยแบ่งตามเขตที่ตั้งของบ้าน ดังรูปที่ 3 จะเห็นว่าข้อมูลมีการกระจายเป็น 5 กลุ่มตามเขตที่ตั้งของบ้าน โดยกลุ่มที่เห็นได้ชัดเจน คือบ้านที่อยู่ในเขต Manhattan ซึ่งข้อมูลมีการกระจายอยู่ในช่วงราคาที่สูงเมื่อเทียบกับบ้านที่อยู่ในเขตอื่น ๆ ส่วนบ้านที่อยู่ในเขต Brooklyn ข้อมูลมีการกระจายอยู่ในช่วงราคาที่ต่ำเมื่อเทียบกับบ้านที่อยู่ในเขตอื่น ๆ

2.4 กำหนดตัวแบบการถดถอยเชิงเส้นและสมการเชิงเส้น

จากการศึกษาข้อมูลเบื้องต้น พบว่าข้อมูลมีตัวแปรอิสระมากกว่า 1 ตัว จึงต้องใช้ตัวแบบการถดถอยเชิงเส้นพหุคูณ ดังนั้นตัวแบบการถดถอยที่กำหนดในการใช้พิจารณาข้อมูลชุดนี้คือ

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 d_{i4} + \beta_5 d_{i5} + \beta_6 d_{i6} + \beta_7 d_{i7} + \beta_8 c_{i8} + \beta_9 c_{i9} + \beta_{10} c_{i10} + \beta_{11} c_{i11} + \varepsilon_i$$

โดยที่ $i = 1, 2, 3, \ldots, 4383$

เมื่อ y_i คือ ราคาบ้านในเมืองนิวยอร์กหลังที่ i

$\beta_0, \beta_1, \ldots, \beta_{11}$ คือ พารามิเตอร์ (สัมประสิทธิ์การถดถอย)

x_{i1} คือ จำนวนห้องนอนของบ้านหลังที่ i

x_{i2} คือ จำนวนห้องน้ำของบ้านหลังที่ i

x_{i3} คือ พื้นที่บ้านรวมที่ดินของบ้านหลังที่ i

d_{i4}, d_{i5}, d_{i6} และ d_{i7} เป็นตัวแปรหุ่นสำหรับการจำแนกประเภทของบ้าน โดยกำหนดค่าดังนี้

$$d_{i4} = \begin{cases} 1 & \text{ถ้าบ้านหลังที่ } i \text{ เป็น Condo} \\ 0 & \text{ถ้าบ้านหลังที่ } i \text{ เป็นบ้านประเภทอื่นๆ} \end{cases}$$

$$d_{i5} = \begin{cases} 1 & \text{ถ้าบ้านหลังที่ } i \text{ เป็น House} \\ 0 & \text{ถ้าบ้านหลังที่ } i \text{ เป็นบ้านประเภทอื่นๆ} \end{cases}$$

$$d_{i6} = \begin{cases} 1 & \text{ถ้าบ้านหลังที่ } i \text{ เป็น Multi-family home} \\ 0 & \text{ถ้าบ้านหลังที่ } i \text{ เป็นบ้านประเภทอื่นๆ} \end{cases}$$

$$d_{i7} = \begin{cases} 1 & \text{ถ้าบ้านหลังที่ } i \text{ เป็น Townhouse} \\ 0 & \text{ถ้าบ้านหลังที่ } i \text{ เป็นบ้านประเภทอื่นๆ} \end{cases}$$

$c_{i8}, c_{i9}, c_{i10}, c_{i11}$ เป็นตัวแปรหุ่นสำหรับการจำแนกเขตที่ตั้งของบ้านในเมืองนิวยอร์ก โดยกำหนดค่าดังนี้

$$c_{i8} = \begin{cases} 1 & \text{ถ้าบ้านหลังที่ } i \text{ อยู่ในเขต Manhattan} \\ 0 & \text{ถ้าบ้านหลังที่ } i \text{ อยู่ในเขตอื่นๆ} \end{cases}$$

$$c_{i9} = \begin{cases} 1 & \text{ถ้าบ้านหลังที่ } i \text{ อยู่ในเขต Queens} \\ 0 & \text{ถ้าบ้านหลังที่ } i \text{ อยู่ในเขตอื่นๆ} \end{cases}$$

$$c_{i10} = \begin{cases} 1 & \text{ถ้าบ้านหลังที่ } i \text{ อยู่ในเขต Staten Island} \\ 0 & \text{ถ้าบ้านหลังที่ } i \text{ เป็นบ้านประเภทอื่นๆ} \end{cases}$$

$$c_{i11} = \begin{cases} 1 & \text{ถ้าบ้านหลังที่ } i \text{ เป็น The Bronx} \\ 0 & \text{ถ้าบ้านหลังที่ } i \text{ เป็นบ้านประเภทอื่นๆ} \end{cases}$$

ε_i คือ ความคลาดเคลื่อนของราคาบ้านในเมืองนิวยอร์กหลังที่ i

และสมการถดถอยของข้อมูลชุดนี้ คือ

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 d_{i4} + \hat{\beta}_5 d_{i5} + \hat{\beta}_6 d_{i6} + \hat{\beta}_7 d_{i7} + \hat{\beta}_8 c_{i8} + \hat{\beta}_9 c_{i9} + \hat{\beta}_{10} c_{i10} + \hat{\beta}_{11} c_{i11}$$

2.5 ประมาณสมการถดถอยที่สอดคล้องกับตัวแบบที่กำหนด

สำหรับการประมาณสมการถดถอยของข้อมูลชุดนี้จะทำโดยการนำข้อมูลมาประมาณค่าสัมประสิทธิ์การถดถอย (β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7, β_8, β_9, β_{10}และ β_{11}) โดยใช้โปรแกรม R ในการคำนวณ และได้ค่าประมาณออกมาดังตารางนี้

	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$		Estimate
(Intercept)	$\hat{\beta}_0$	-	1,267,000
BEDS	$\hat{\beta}_1$	-	120,000
BATH	$\hat{\beta}_2$		502,200
PROPERTYSQFT	$\hat{\beta}_3$		507.10
TYPE_Condo	$\hat{\beta}_4$		1,019,000
TYPE_House	$\hat{\beta}_5$		965,000
TYPE_Multi-family home	$\hat{\beta}_6$		121,600
TYPE_Townhouse	$\hat{\beta}_7$		3,281,000
COUNTY_Manhattan	$\hat{\beta}_8$		2,641,000
COUNTY_Queens	$\hat{\beta}_9$	-	88,700
COUNTY_Staten Island	$\hat{\beta}_{10}$	-	714,200
COUNTY_The Bronx	$\hat{\beta}_{11}$	-	184,400

ตารางที่ 4 ตารางค่าประมาณสัมประสิทธิ์การถดถอย

สมการถดถอยที่ประมาณได้คือ

$$\hat{y} = -1{,}267{,}000 - 120{,}000x_1 + 502{,}200x_2 + 507.10x_3 + 1{,}019{,}000d_4 + 965{,}000d_5 + 121{,}600d_6 + 3{,}281{,}000d_7 + 2{,}641{,}000c_8 - 88{,}700c_9 - 714{,}200c_{10} - 184{,}400c_{11}$$

2.6 ตรวจสอบสัมประสิทธิ์ของสมการถดถอยที่ประมาณได้

การตรวจสอบในที่นี้จะเป็นการทดสอบความมีนัยสำคัญของสมการถดถอย โดยทดสอบว่าจำนวนห้องนอน จำนวนห้องน้ำ พื้นที่บ้านรวมที่ดิน ประเภทของบ้าน และเขตของบ้านที่ตั้งในเมืองนิวยอร์ก สามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้อย่างมีนัยสำคัญทางสถิติหรือไม่ ซึ่งใช้โปรแกรม R ในการทดสอบ ได้ค่าดังตารางต่อไปนี้

	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$	t-test	P-value
(Intercept)	$\hat{\beta}_0$	-8.797	0.00000
BEDS	$\hat{\beta}_1$	-3.56	0.00037
BATH	$\hat{\beta}_2$	11.668	0.00000
PROPERTYSQFT	$\hat{\beta}_3$	21.145	0.00000
TYPE_Condo	$\hat{\beta}_4$	6.992	0.00000
TYPE_House	$\hat{\beta}_5$	6.031	0.00000
TYPE_Multi-family home	$\hat{\beta}_6$	0.624	0.53291
TYPE_Townhouse	$\hat{\beta}_7$	14.178	0.00000
COUNTY_Manhattan	$\hat{\beta}_8$	17.399	0.00000
COUNTY_Queens	$\hat{\beta}_9$	-0.617	0.53731
COUNTY_Staten Island	$\hat{\beta}_{10}$	-3.591	0.00033
COUNTY_The Bronx	$\hat{\beta}_{11}$	-0.987	0.32349

	F-test	P-value
ALL Coefficients	221.8	0.00000

ตารางที่ 5 ตารางค่า P-value ของสัมประสิทธิ์การถดถอย

โดยเริ่มจากการทดสอบสมการถดถอยทั้งสมการ ที่ระดับนัยสำคัญ 0.05 ก่อน ดังนี้

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$

H_a : มี $\beta_j \neq 0$ อย่างน้อยหนึ่งค่า

จากตารางที่ 5 ค่าสถิติทดสอบ F = 221.8 และค่า P-value = 0.000 ซึ่งมีค่าน้อยกว่า 0.05 จึงปฏิเสธสมมติฐานหลัก หมายความว่า จำนวนห้องนอน จำนวนห้องน้ำ พื้นที่บ้านรวมที่ดิน ประเภทของบ้าน และเขตที่ตั้งของบ้านในเมืองนิวยอร์ก มีอย่างน้อยหนึ่งตัวแปรที่สามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้ที่ระดับนัยสำคัญ 0.05

จากการทดสอบสมมติฐานข้างต้น พบว่ามีตัวแปรอิสระอย่างน้อยหนึ่งตัวแปรที่สามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้ จึงต้องทดสอบสัมประสิทธิ์การถดถอยของตัวแปรอิสระแต่ละตัว เพื่อหาว่ามีตัวแปรใดบ้างที่สามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้ ดังนั้นจึงทดสอบสัมประสิทธิ์การถดถอยทีละตัว โดยเริ่มจากการทดสอบสมมติฐานเกี่ยวกับ β_1 ซึ่งเป็นสัมประสิทธิ์การถดถอยของจำนวนห้องนอน

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

จากตารางที่ 5 ค่าสถิติทดสอบ t = |-3.560| และค่า P-value = 0.000 ซึ่งมีค่าน้อยกว่า 0.05 จึงปฏิเสธสมมติฐานหลัก หมายความว่า จำนวนห้องนอนสามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้

ต่อไปจะเป็นการทดสอบสมมติฐานเกี่ยวกับ β_2 ซึ่งเป็นสัมประสิทธิ์การถดถอยของจำนวนห้องน้ำ

$H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

จากตารางที่ 5 ค่าสถิติทดสอบ t = |11.668| และค่า P-value = 0.000 ซึ่งมีค่าน้อยกว่า 0.05 จึงปฏิเสธสมมติฐานหลัก หมายความว่า จำนวนห้องน้ำสามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้

ต่อไปจะเป็นการทดสอบสมมติฐานเกี่ยวกับ β_3 ซึ่งเป็นสัมประสิทธิ์การถดถอยของพื้นที่บ้านรวมที่ดิน

$H_0: \beta_3 = 0$

$H_a: \beta_3 \neq 0$

จากตารางที่ 5 ค่าสถิติทดสอบ t = |21.145| และค่า P-value = 0.000 ซึ่งมีค่าน้อยกว่า 0.05 จึงปฏิเสธสมมติฐานหลัก หมายความว่า พื้นที่บ้านรวมที่ดินสามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้

สำหรับการทดสอบสมมติฐานเกี่ยวกับ β_4, β_5, β_6 และ β_7 ซึ่งเป็นสัมประสิทธิ์การถดถอยของตัวแปรหุ่น d_4, d_5, d_6 และ d_7 ซึ่งเป็นตัวแทนของตัวแปรอิสระเชิงคุณภาพตัวเดียวกันคือประเภทของบ้าน การทดสอบ

สมมติฐานจะไม่สามารถแยกทดสอบ β_4 β_5 β_6 และ β_7 ได้ จึงต้องทดสอบพร้อมกันทั้งสี่ค่า โดยใช้การทดสอบ partial F-test และ P-value ซึ่งใช้โปรแกรม R ในการทดสอบ ได้ค่าดังตารางต่อไปนี้

	partial F-test	P-value
TYPE	100.26	0.000
COUNTY	118.05	0.000

ตารางที่ 6 ตารางค่า P-value ของสัมประสิทธิ์การถดถอยของตัวแปรหุ่น

H_0 : $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

H_a : มี $\beta_j \neq 0$ อย่างน้อยหนึ่งค่า

จากตารางที่ 6 ค่าสถิติทดสอบ partial F = 100.26 และค่า P-value = 0.000 ซึ่งค่า P-value ที่ได้น้อยกว่า 0.05 จึงปฏิเสธสมมติฐานหลัก หมายความว่า บ้านที่จัดอยู่ในประเภทที่แตกต่างกันจะมีราคาบ้านแตกต่างกันที่ระดับนัยสำคัญ 0.05

และการทดสอบสมมติฐานเกี่ยวกับ β_8, β_9, β_{10} และ β_{11} ซึ่งเป็นสัมประสิทธิ์การถดถอยของตัวแปรหุ่น d_8, d_9, d_{10} และ d_{11} ซึ่งเป็นตัวแทนของตัวแปรอิสระเชิงคุณภาพตัวเดียวกันคือเขตที่ตั้งของบ้านในเมืองนิวยอร์ก จึงต้องทดสอบ β_8, β_9, β_{10} และ β_{11} พร้อมกันทั้งสี่ค่า โดยใช้การทดสอบ partial F-test และ P-value ดังนี้

H_0 : $\beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$

H_a : มี $\beta_j \neq 0$ อย่างน้อยหนึ่งค่า

จากตารางที่ 6 ค่าสถิติทดสอบ partial F = 118.05 และค่า P-value = 0.000 ซึ่งค่า P-value ที่ได้น้อยกว่า 0.05 จึงปฏิเสธสมมติฐานหลัก หมายความว่า บ้านที่ตั้งอยู่ในเขตที่แตกต่างกันจะมีราคาบ้านที่แตกต่างกันที่ระดับนัยสำคัญ 0.05

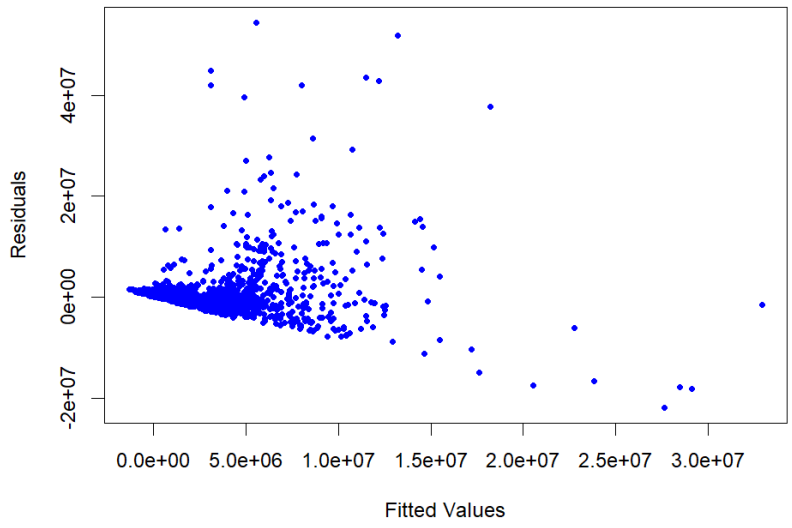
2.7 ตรวจสอบข้อสมมติของตัวแบบการถดถอย

ขั้นตอนนี้จะเป็นการตรวจสอบความเหมาะสมของตัวแบบในการนำไปใช้ทำนายค่าตัวแปรตาม Y หรือการประมาณค่าเฉลี่ยของราคาบ้าน โดยการตรวจสอบข้อสมมติที่เป็นไปได้ของตัวแบบที่ถูกกำหนดขึ้นว่ามีความถูกต้องหรือไม่ จะเป็นการตรวจสอบโดยใช้ส่วนเหลือ (residual : $\hat{\varepsilon}$) ของตัวแบบ ซึ่งวิธีการตรวจสอบข้อสมมติจะมีทั้งการใช้แผนภาพส่วนเหลือชนิดต่างๆ เช่น ฮิสโทแกรม แผนภาพการกระจาย แผนภาพความน่าจะเป็นปกติ เป็นต้น และการใช้การทดสอบเชิงสถิติต่างๆ อย่าง Anderson-Darling test, Durbin Watson และข้อสมมติที่ต้องตรวจสอบนั้น ได้แก่

1. ความสัมพันธ์ระหว่างตัวแปรตาม y กับตัวแปรอิสระ ต้องเป็นความสัมพันธ์เชิงเส้น
2. ความคลาดเคลื่อน ε มีการแจกแจงปกติมีค่าเฉลี่ยเท่ากับ 0
3. ความคลาดเคลื่อน ε มีความแปรปรวนคงที่คือ σ^2
4. ความคลาดเคลื่อนไม่มีความสัมพันธ์กัน
5. ตัวแปรอิสระในตัวแบบไม่มีความสัมพันธ์กัน

2.7.1 การตรวจสอบความสัมพันธ์เชิงเส้นของตัวแบบการถดถอย

สำหรับข้อสมมติที่ว่าตัวแปรตามและตัวแปรอิสระต้องมีความสัมพันธ์กันในเชิงเส้น จะตรวจสอบโดยการใช้แผนภาพการกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และค่าประมาณของตัวแปรตาม ($\hat{y}$) โดยมีวิธีการพิจารณาคือ หากแผนภาพที่ได้มีลักษณะกระจายแบบสุ่มรอบศูนย์ แสดงว่า ตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกัน โดยที่นี้จะใช้โปรแกรม R ในการสร้างแผนภาพ และได้แผนภาพการกระจายออกมาดังนี้

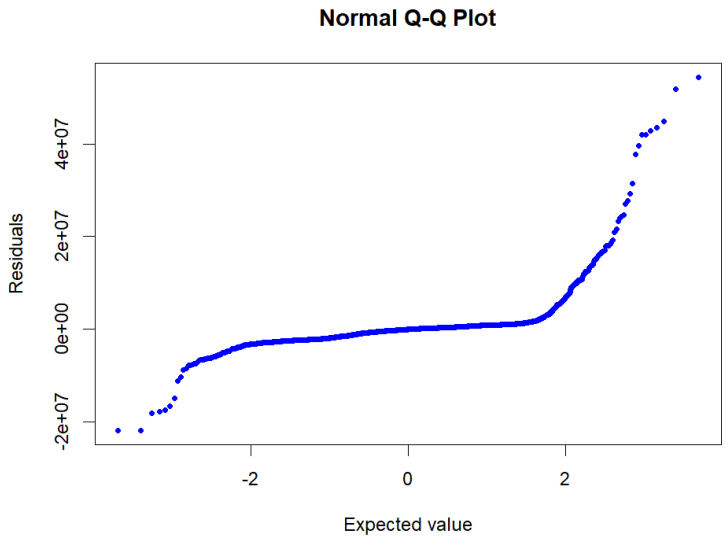


รูปที่ 4 แผนภาพการกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และค่าประมาณของตัวแปรตาม ($\hat{y}$)

จากรูปที่ 4 พบว่าจุดของข้อมูลบนแผนภาพการกระจายไม่ได้มีการกระจายสุ่มรอบศูนย์ ตามสมบัติของส่วนเหลือ ($\hat{\varepsilon}$) นั้นแสดงให้เห็นว่า ตัวแปรตามและตัวแปรอิสระไม่มีความสัมพันธ์เชิงเส้นกัน และสมการถดถอยเชิงเส้นที่ประมาณได้อาจจะไม่มีความเหมาะสมกับข้อมูลชุดนี้

2.7.2 การตรวจสอบการแจกแจงปกติของความคลาดเคลื่อน

สำหรับข้อสมมติของตัวแบบที่ว่า ความคลาดเคลื่อน (ε) ต้องมีการแจกแจงปกติ จะตรวจสอบโดยการใช้แผนภาพความน่าจะเป็นปกติ โดยใช้โปรแกรม R ในการสร้างแผนภาพ และได้แผนภาพการกระจายออกมาดังนี้



รูปที่ 5 แผนภาพความน่าจะเป็นปกติ

จากรูปที่ 5 พบว่าข้อมูลส่วนใหญ่มีการกระจายตัวอยู่ตรงกลางบริเวณค่าเฉลี่ยของความคลาดเคลื่อน และข้อมูลส่วนน้อยจะมีการกระจายในตำแหน่งที่สูงหรือต่ำกว่าค่าเฉลี่ย จึงถือว่าแผนภาพความคลาดเคลื่อนมีการแจกแจงแบบหางยาว (long-tailed distribution) แสดงว่าความคลาดเคลื่อนไม่มีการแจกแจงปกติ

นอกจากการใช้แผนภาพในการตรวจสอบข้อสมมติเกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนแล้ว ยังสามารถใช้การทดสอบเชิงสถิติในการตรวจสอบข้อสมมตินี้ได้ โดยจะใช้วิธีการทดสอบเชิงสถิติอย่าง Anderson-Darling test ที่ใช้โปรแกรม R ในการหาค่า ซึ่งได้ค่าสถิติทดสอบ Anderson-Darling เท่ากับ 506.65 และค่า P-value เท่ากับ 0.000 ดังรูป ช. (ภาคผนวก ค.)

สมมติฐาน:

H_0: ความคลาดเคลื่อนมีการแจกแจงปกติ

H_a: ความคลาดเคลื่อนไม่มีการแจกแจงปกติ

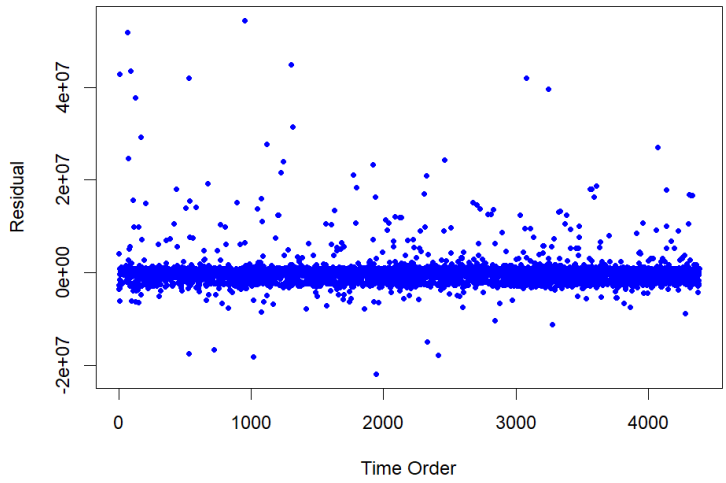
เนื่องจาก ค่า P-value ที่ได้น้อยกว่าระดับนัยสำคัญ 0.05 จึงปฏิเสธสมมติฐานหลัก สรุปได้ว่าความคลาดเคลื่อนไม่มีการแจกแจงปกติ

2.7.3 การตรวจสอบความคงที่ของความแปรปรวนของความคลาดเคลื่อน

จากข้อสมมติที่ว่าความคลาดเคลื่อน (ε) ต้องมีความแปรปรวนคงที่ โดยในการตรวจสอบข้อสมมตินี้จะใช้แผนภาพการกระจายในรูปที่ 4 ซึ่งเห็นว่าข้อมูลมีการกระจายเป็นรูปกรวย นั่นหมายความว่า ราคาบ้านที่ประมาณได้ ($\hat{y}$) ในราคาสูงนั้นมีการกระจายของส่วนเหลือ ($\hat{\varepsilon}$) มากกว่าราคาบ้านที่ประมาณได้ในราคาต่ำ ดังนั้นแผนภาพนี้จึงชี้ให้เห็นว่าความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ หรือภาวะความแปรปรวนต่างกัน (Heteroscedasticity)

2.7.4 การตรวจสอบความเป็นอิสระกันของความคลาดเคลื่อน

สำหรับข้อสมมติที่ว่าความคลาดเคลื่อน (ε) ต้องเป็นอิสระกัน จะตรวจสอบโดยใช้แผนภาพการกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และลำดับของการเก็บข้อมูล โดยถ้าแผนภาพการกระจายมีลักษณะการกระจายรอบ ๆ ศูนย์ในรูปแบบสุ่มหรือใกล้เคียง แสดงว่าความคลาดเคลื่อนเป็นอิสระกัน โดยที่นี้จะใช้โปรแกรม R ในการสร้างแผนภาพ และได้แผนภาพการกระจายออกมาดังนี้



รูปที่ 6 แผนภาพการกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และลำดับของการเก็บข้อมูล

จากรูปที่ 6 การกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และลำดับของการเก็บข้อมูล มีการกระจายรอบๆค่าศูนย์ในรูปแบบสุ่ม แสดงว่าความคลาดเคลื่อนเป็นอิสระกัน

นอกจากพิจารณาจากแผนภาพการกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และลำดับของการเก็บข้อมูลแล้ว ยังสามารถใช้การทดสอบเชิงสถิติที่เรียกว่าการทดสอบ Durbin-Watson (Durbin-Watson test) ซึ่งได้ค่าสถิติทดสอบ Durbin-Watson เท่ากับ 2.0462 และค่า P-value เท่ากับ 0.9367 ดังรูป ฌ. (ภาคผนวก ค.)

สมมติฐาน:

H_0: ความคลาดเคลื่อนไม่มีความสัมพันธ์กัน

H_a: ความคลาดเคลื่อนมีความสัมพันธ์กันในทางบวก

เนื่องจาก ค่า P-value ที่ได้มากกว่าระดับนัยสำคัญ 0.05 จึงยอมรับสมมติฐานหลัก สรุปได้ว่าความคลาดเคลื่อนไม่มีความสัมพันธ์กันหรือความคลาดเคลื่อนเป็นอิสระกัน

2.7.5 การตรวจสอบความเป็นอิสระกันของตัวแปรอิสระในตัวแบบ

สำหรับตัวแบบการถดถอยที่ประกอบด้วยตัวแปรอิสระมากกว่าหนึ่งตัว อาจเกิดปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กันเอง ที่เรียกว่าการเกิดความสัมพันธ์เชิงเส้นแบบพหุ (multicollinearity) ซึ่งความรุนแรงของปัญหาที่เกิดขึ้น ขึ้นอยู่กับขนาดของความสัมพันธ์ระหว่างตัวแปรอิสระในตัวแบบ

การตรวจสอบความเป็นอิสระกันของตัวแปรอิสระในตัวแบบการถดถอยมีอยู่ด้วยกันหลายวิธี โดยเราจะใช้วิธีพิจารณาจากค่า Variance inflation factor (VIF) ซึ่งค่า VIF เป็นค่าวัดขนาดของความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรอิสระตัวที่เหลือในสมการถดถอย เมื่อค่า VIF มีค่ามากจะชี้ให้เห็นว่า x_j อาจมีความสัมพันธ์กับตัวแปรอิสระตัวอื่น ๆ สูง จึงเกิดปัญหาความสัมพันธ์เชิงเส้นแบบพหุขึ้น โดยเกณฑ์ของการใช้ค่า VIF คือ หาก VIF_j มีค่ามากกว่า 10 ชี้ให้เห็นว่าเกิดปัญหาความสัมพันธ์เชิงเส้นแบบพหุค่อนข้างรุนแรง โดยได้ผลการคำนวณจากโปรแกรม R ดังนี้

	VIF
BEDS	3.143773
BATH	2.867375
PROPERTYSQFT	1.342201
TYPE	2.173051
COUNTY	1.462001

ตารางที่ 7 ตารางค่า Variance inflation factor (VIF)

จากตารางที่ 7 พิจารณาค่า VIF ของตัวแปรอิสระแต่ละตัว พบว่าค่า VIF ของ BEDS BATH PROPERTYSQFT TYPE และ COUNTY เท่ากับ 3.14 2.87 1.34 2.17 และ 1.46 ตามลำดับ ชี้ให้เห็นว่ามีความสัมพันธ์เชิงเส้นแบบพหุค่อนข้างต่ำ แสดงว่าตัวแปรอิสระแต่ละตัวมีความเป็นอิสระกัน

2.8 การแปลงรูปข้อมูล

จากการตรวจสอบข้อสมมติของตัวแบบการถดถอยทั้ง 5 ข้อสมมติ พบว่าตัวแปรตามและตัวแปรอิสระไม่ได้มีความสัมพันธ์กันในเชิงเส้น ความคลาดเคลื่อนไม่มีการแจกแจงปกติ และความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ ซึ่งการแก้ปัญหานี้จะใช้วิธีการแปลงค่าของตัวแปรตาม เนื่องจากต้องการเปลี่ยนรูปร่างและการกระจายของการแจกแจงของตัวแปรตาม โดยแปลงรูป y ให้เป็น y' โดยที่ $y' = \log_{10} y$ ซึ่งได้ค่า y' แสดงในคอลัมน์สุดท้ายของตารางที่ 8 ดังนี้

i	PRICE (y)	BEDS	BATH	PROPERTYSQFT	TYPE	COUNTY	$y' = \log_{10} y$
1	315,000	2	2	1400.00	Condo	Manhattan	5.498311
2	19,500,000	7	10	17545.00	Condo	Manhattan	7.290035
3	69,000	3	1	445.00	Condo	Manhattan	4.838849
4	899,500	2	2	2184.21	Condo	Manhattan	5.954001
5	260,000	4	2	2015.00	House	Staten Island	5.414973
6	690,000	5	2	4004.00	House	Brooklyn	5.838849
7	55,000,000	7	2	14175.00	Townhouse	Manhattan	7.740363
8	16,800,000	8	16	33000.00	House	Staten Island	7.225309
9	265,000	1	1	750.00	Co-op	The Bronx	5.423246
10	440,000	2	1	978.00	Co-op	Brooklyn	5.643453

ตารางที่ 8 การใช้ลอการิทึมฐานสิบแปลงรูปตัวแปรตาม (y)

เพื่อตรวจสอบความเหมาะสมของการแปลงรูป $y' = \log_{10} y$ จึงได้ประมาณตัวแบบการถดถอยเชิงเส้นพหุคูณโดยใช้ตัวแปรตามคือ y' ได้ค่าประมาณสัมประสิทธิ์การถดถอยออกมาดังตารางนี้

	$\hat{\beta}_j$	Estimate
(Intercept)	$\hat{\beta}_0$	5.47100
BEDS	$\hat{\beta}_1$	-0.00749
BATH	$\hat{\beta}_2$	0.06769
PROPERTYSQFT	$\hat{\beta}_3$	0.00003
TYPE_Condo	$\hat{\beta}_4$	0.28300
TYPE_House	$\hat{\beta}_5$	0.38530
TYPE_Multi-family home	$\hat{\beta}_6$	0.38070
TYPE_Townhouse	$\hat{\beta}_7$	0.53910
COUNTY_Manhattan	$\hat{\beta}_8$	0.40080
COUNTY_Queens	$\hat{\beta}_9$	-0.08578
COUNTY_Staten Island	$\hat{\beta}_{10}$	-0.18870
COUNTY_The Bronx	$\hat{\beta}_{11}$	-0.19270

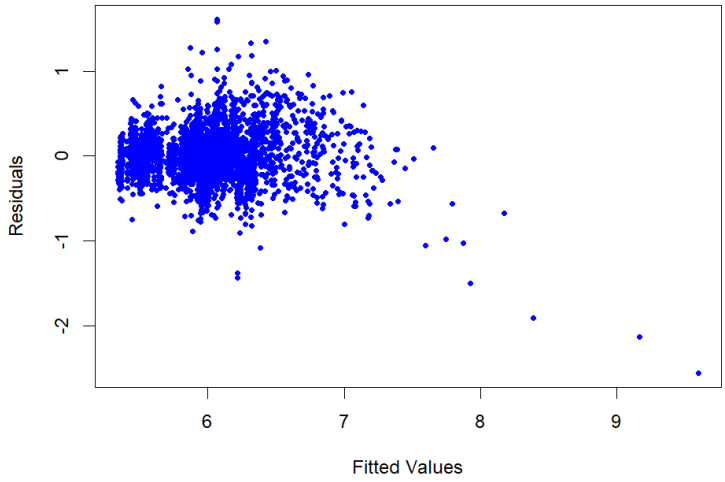
ตารางที่ 9 ตารางค่าประมาณสัมประสิทธิ์การถดถอย

สมการถดถอยที่ประมาณได้คือ

$$\hat{y}' = 5.47100 - 0.00749x_1 + 0.06769x_2 + 0.00003x_3 + 0.28300d_4 + 0.38530d_5 + 0.38070d_6 + 0.53910d_7 + 0.40080c_8 - 0.08578c_9 - 0.18870c_{10} - 0.19270c_{11}$$

2.8.1 ตรวจสอบข้อสมมติของตัวแบบการถดถอย

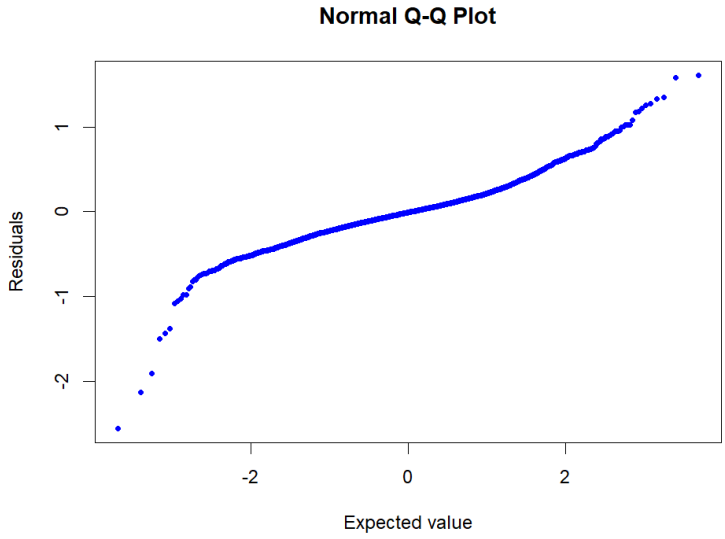
(1) การตรวจสอบความสัมพันธ์เชิงเส้นของตัวแบบการถดถอย



รูปที่ 7 แผนภาพการกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และค่าประมาณของตัวแปรตาม ($\hat{y}'$)

จากรูปที่ 7 พบว่าจุดของข้อมูลบนแผนภาพการกระจายมีการกระจายสุ่มรอบศูนย์ ตามสมบัติของส่วนเหลือ ($\hat{\varepsilon}$) นั้นแสดงให้เห็นว่า ตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกัน

(2) การตรวจสอบการแจกแจงปกติของความคลาดเคลื่อน



รูปที่ 8 แผนภาพความน่าจะเป็นปกติ

จากรูปที่ 8 พบว่าข้อมูลส่วนใหญ่มีการกระจายตัวอยู่ตรงกลางบริเวณค่าเฉลี่ยของความคลาดเคลื่อน และข้อมูลส่วนน้อยจะมีการกระจายในตำแหน่งที่สูงหรือต่ำกว่าค่าเฉลี่ย จึงถือว่าแผนภาพความคลาดเคลื่อนมีการแจกแจงแบบหางยาว (long-tailed distribution) แสดงว่าความคลาดเคลื่อนไม่มีการแจกแจงปกติ

จากการคำนวณโดยใช้โปรแกรม R ได้ค่าสถิติทดสอบ Anderson-Darling เท่ากับ 36.798 และค่า P-value เท่ากับ 0.000 ดังรูป ฎ. (ภาคผนวก ค.)

สมมติฐาน:

H_0: ความคลาดเคลื่อนมีการแจกแจงปกติ

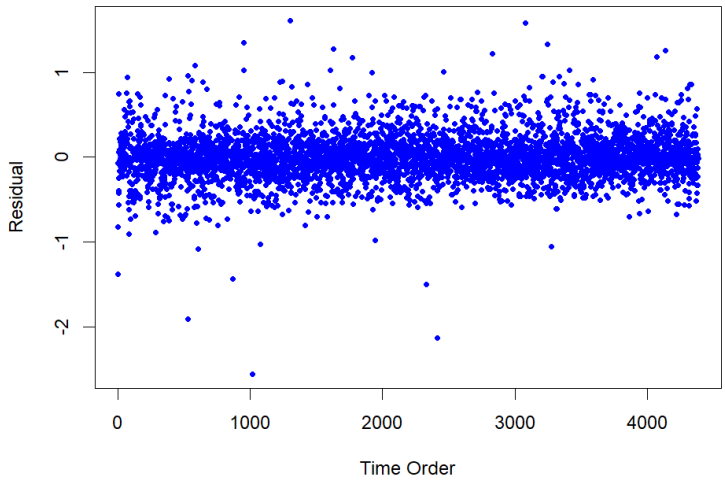
H_a: ความคลาดเคลื่อนไม่มีการแจกแจงปกติ

เนื่องจาก ค่า P-value ที่ได้น้อยกว่าระดับนัยสำคัญ 0.05 จึงปฏิเสธสมมติฐานหลัก สรุปได้ว่าความคลาดเคลื่อนไม่มีการแจกแจงปกติ

(3) การตรวจสอบความคงที่ของความแปรปรวนของความคลาดเคลื่อน

จากแผนภาพการกระจายในรูปที่ 7 ซึ่งเห็นว่าข้อมูลมีการกระจายสุ่มรอบศูนย์ แสดงว่าความคลาดเคลื่อนมีความแปรปรวนคงที่ หรือภาวะความแปรปรวนเท่ากัน (Homoscedasticity)

(4) การตรวจสอบความเป็นอิสระกันของความคลาดเคลื่อน



รูปที่ 9 แผนภาพการกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และลำดับของการเก็บข้อมูล

จากรูปที่ 9 การกระจายระหว่างส่วนเหลือ ($\hat{\varepsilon}$) และลำดับของการเก็บข้อมูล มีการกระจายรอบ ๆ ค่าศูนย์ในรูปแบบสุ่ม แสดงว่าความคลาดเคลื่อนเป็นอิสระกัน

การทดสอบ Durbin-Watson ได้ค่าสถิติทดสอบ Durbin-Watson เท่ากับ 1.961 และค่า P-value เท่ากับ 0.0979 ดังรูป ฎ. (ภาคผนวก ค.)

สมมติฐาน:

H_0: ความคลาดเคลื่อนไม่มีความสัมพันธ์กัน

H_a: ความคลาดเคลื่อนมีความสัมพันธ์กันในทางบวก

เนื่องจาก ค่า P-value ที่ได้มากกว่าระดับนัยสำคัญ 0.05 จึงยอมรับสมมติฐานหลัก สรุปได้ว่าความคลาดเคลื่อนไม่มีความสัมพันธ์กันหรือความคลาดเคลื่อนเป็นอิสระกัน

(5) การตรวจสอบความเป็นอิสระกันของตัวแปรอิสระในตัวแบบ

	VIF
BEDS	3.143773
BATH	2.867375
PROPERTYSQFT	1.342201
TYPE	2.173051
COUNTY	1.462001

ตารางที่ 10 ตารางค่า Variance inflation factor (VIF)

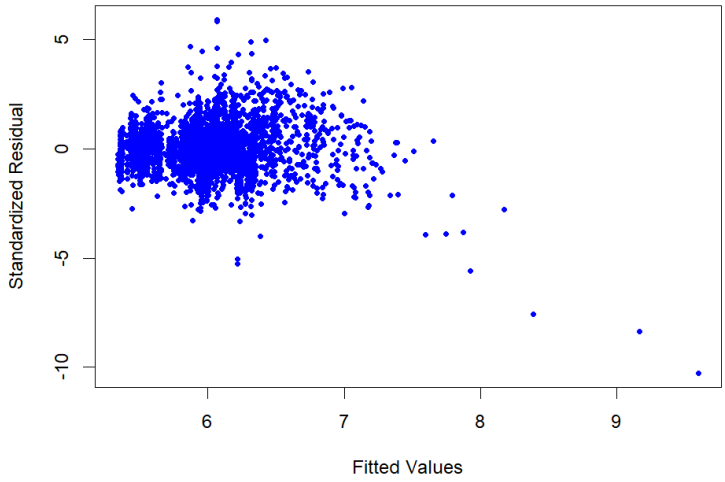
จากตารางที่ 10 พิจารณาค่า VIF ของตัวแปรอิสระแต่ละตัว พบว่าค่า VIF ของ BEDS BATH PROPERTYSQFT TYPE และ COUNTY เท่ากับ 3.14 2.87 1.34 2.17 และ 1.46 ตามลำดับ ชี้ให้เห็นว่ามีความสัมพันธ์เชิงเส้นแบบพหุค่อนข้างต่ำ แสดงว่าตัวแปรอิสระแต่ละตัวมีความเป็นอิสระกัน

2.9 ตรวจสอบค่านอกเกณฑ์และค่าที่มีอิทธิพล

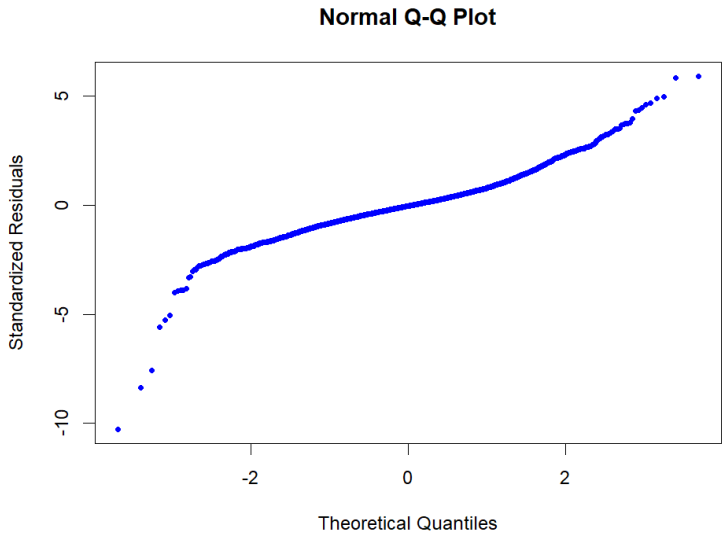
2.9.1 ตรวจสอบค่านอกเกณฑ์

ในการวิเคราะห์การถดถอยนอกจากการตรวจสอบข้อสมมติของตัวแบบการถดถอยแล้ว ยังควรตรวจสอบด้วยว่าในชุดข้อมูลนั้นมีค่าสังเกตที่เป็นค่านอกเกณฑ์ด้วย เนื่องจากค่านอกเกณฑ์จะส่งผลกระทบต่อสมการถดถอยและการวิเคราะห์การถดถอย ในที่นี้จะใช้โปรแกรม R ในการสร้างแผนภาพการกระจายของส่วนเหลือมาตรฐาน (standardized residual) ได้ดังรูปที่ 10 และแผนภาพความน่าจะเป็นปกติของส่วนเหลือมาตรฐานได้ดังรูปที่ 11 ในการพิจารณาค่านอกเกณฑ์

จากแผนภาพในรูปที่ 10 และรูปที่ 11 จะเห็นได้ว่ามีค่าสังเกตที่มีค่าส่วนเหลือมาตรฐานมากกว่า 3 หรือน้อยกว่า -3 เป็นจำนวนมาก ซึ่งมีทั้งหมด 48 ค่าสังเกต จึงถือว่าค่าเหล่านี้คือค่านอกเกณฑ์ แต่เนื่องจากค่านอกเกณฑ์มีหลายค่าซึ่งจะส่งผลให้การวิเคราะห์มีความยุ่งยาก จึงจำเป็นต้องตรวจสอบว่าค่านอกเกณฑ์เหล่านี้เกิดขึ้นจากสาเหตุอะไรได้บ้าง เช่น ความผิดพลาดในการบันทึก การวัดค่าผิด เครื่องมือวัดมีสภาพชำรุด หรือตัวแบบการถดถอยขาดตัวแปรอิสระที่สำคัญไป แต่เนื่องจากข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลที่มีผู้รวบรวมไว้อยู่แล้ว จึงไม่สามารถตรวจสอบและแก้ผลกระทบต่อการวิเคราะห์การถดถอยที่เกิดขึ้น หรือไม่สามารถตัดค่าสังเกตนี้ออกไปจากชุดข้อมูลได้



รูปที่ 10 แผนภาพการกระจายระหว่างส่วนเหลือมาตรฐานและค่าประมาณของตัวแปรตาม ($\hat{y}'$)



รูปที่ 11 แผนภาพความน่าจะเป็นปกติของส่วนเหลือมาตรฐาน

2.9.2 ตรวจสอบค่าที่มีอิทธิพล

จากแผนภาพในรูปที่ 10 จะเห็นว่ากลุ่มจุดที่อยู่มุมล่างขวามือเป็นจุดที่มีค่าประมาณของตัวแปรตาม ($\hat{y}'$) มากที่สุดและแตกต่างจากค่าสังเกตอื่นๆมาก (ค่าสังเกตที่ 1014, 2414, 528, 2332) เรียกจุดกลุ่มข้อมูลหรือค่าสังเกตนี้ว่าจุด Leverage ที่ถือเป็นค่านอกเกณฑ์ในค่าประมาณของตัวแปรตาม ซึ่งค่า Leverage ของค่าสังเกตจะหาได้โดยการใช้โปรแกรม R ในการคำนวณ และค่า Leverage ของกลุ่มจุดที่อยู่มุมล่างขวามือเมื่อเทียบกับเกณฑ์มาตรฐานในการพิจารณา คือ 2p/n = 0.005476 และ 3p/n = 0.008214 แล้วมีค่ามากกว่า ค่าสังเกตเหล่านี้จึงมีอิทธิพลต่อค่าประมาณสัมประสิทธิ์การถดถอย นอกจากการใช้ค่า Leverage ในการตรวจสอบค่าที่มีอิทธิพลแล้ว ยังสามารถใช้ค่า Cook's distance (D) และ DFFITS ที่ใช้โปรแกรม R ในการคำนวณได้อีกด้วย และได้ผลดังตารางนี้

i	PRICE	BEDS	BATH	PROPERTY	TYPE	COUNTY	SRES	h	D	DFFITS
1014	11000000	50	50	22035	Multi-family	Manhattai	-10.2739	0.164788	1.735462	-4.61909
2414	10700000	42	43	27152	Multi-family	Manhattai	-8.37011	0.121522	0.807615	-3.13799
528	3000000	3	32	11760	Multi-family	Manhattai	-7.57324	0.140419	0.780768	-3.08085
2332	2700000	24	24	18936	Multi-family	Manhattai	-5.5945	0.033714	0.091002	-1.04864

ตารางที่ 11 ตารางค่าส่วนเหลือมาตรฐาน ค่า Leverage, Cook's distance และ DFFITS

ดังนั้น จึงสรุปได้ว่ากลุ่มจุดที่อยู่มุมล่างขวามือในแผนภาพรูปที่ 7 นี้เป็นค่านอกเกณฑ์และเป็นค่าที่มีอิทธิพลด้วย และการจะแก้ไขปัญหาด้วยการตัดค่าสังเกตที่มีอิทธิพลนี้ออกไปนั้นไม่สามารถทำได้ เนื่องจากค่าสังเกตนี้เป็นสิ่งที่สำคัญในการการศึกษาในครั้งนี้

2.10 การคัดเลือกตัวแปรอิสระและตัวแบบที่เหมาะสม

จากหัวข้อ 2.6 ที่ผ่านมาได้ใช้การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอยในการพิจารณาว่าตัวแปรอิสระใดที่มีความสำคัญและสามารถอธิบายความแปรผันของตัวแปรตามได้ โดยใช้การทดสอบเอฟสำหรับการทดสอบสัมประสิทธิ์การถดถอยทั้งตัวแบบ ใช้การทดสอบเอฟบางส่วนและการทดสอบทีสำหรับการทดสอบสัมประสิทธิ์การถดถอยบางตัวในตัวแบบ

นอกจากการทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอยแล้ว ยังมีวิธีอื่น ๆ ที่นิยมใช้ในการคัดเลือกตัวแปรอิสระที่สำคัญ ซึ่งเรียกว่า วิธีการคัดเลือกตัวแปร (variable-selection techniques) โดยจะใช้ 3 วิธี ได้แก่

1. การเลือกแบบไปข้างหน้า (Forward selection procedure)

2. การกำจัดแบบถอยหลัง (Backward elimination procedure)

3. การถดถอยทีละขั้น (Stepwise regression procedure)

2.10.1 การเลือกแบบไปข้างหน้า

วิธีนี้จะเริ่มจากตัวแบบที่ไม่มีตัวแปรอิสระตัวใดอยู่เลย จากนั้นค่อย ๆ เพิ่มตัวแปรอิสระเข้าไปในตัวแบบทีละตัว โดยเพิ่มตัวแปรอิสระที่มีความสำคัญหรืออธิบายความแปรผันของตัวแปรตามได้มากที่สุดเข้าไปเป็นตัวแรก

จากการใช้โปรแกรม R ทำวิธี Forward selection ในการคัดเลือกตัวแบบที่เหมาะสมได้ผลดังรูป ฒ. (ภาคผนวก ค.) ได้ตัวแบบการถดถอยที่ประกอบด้วยตัวแปรอิสระจำนวน 5 ตัวแปร ได้แก่ BEDS, BATH, PROPERTYSQFT, TYPE และ COUNTY ได้ค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ดังรูป ณ. (ภาคผนวก ค.) ดังนั้นสมการถดถอยที่ประมาณได้ คือ

$$\widehat{\log_{10} y} = 5.47133 - 0.00749x_1 + 0.06769x_2 + 0.00003x_3 + 0.28296d_4 + 0.38534d_5 + 0.38068d_6 + 0.53909d_7 + 0.40082c_8 - 0.08578c_9 - 0.18870c_{10} - 0.19267c_{11}$$

2.10.2 การกำจัดแบบถอยหลัง

วิธีนี้มีขั้นตอนตรงข้ามกับวิธีการเลือกแบบไปข้างหน้า โดยเริ่มจากตัวแบบที่ประกอบด้วยตัวแปรอิสระทุกตัวแปรก่อน จากนั้นจึงค่อยพิจารณาตัดตัวแปรที่ไม่สามารถอธิบายความแปรผันของตัวแปรตามหรืออธิบายได้น้อยออกทีละตัว

จากการใช้โปรแกรม R ทำวิธี Backward elimination ในการคัดเลือกตัวแบบที่เหมาะสมได้ผลดังรูป ด. (ภาคผนวก ค.) ได้ตัวแบบการถดถอยที่ประกอบด้วยตัวแปรอิสระจำนวน 5 ตัวแปร ได้แก่ BEDS, BATH, PROPERTYSQFT, TYPE และ COUNTY ได้ค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ดังรูป ต. (ภาคผนวก ค.) ดังนั้นสมการถดถอยที่ประมาณได้ คือ

$$\widehat{\log_{10} y} = 5.47133 - 0.00749x_1 + 0.06769x_2 + 0.00003x_3 + 0.28296d_4 + 0.38534d_5 + 0.38068d_6 + 0.53909d_7 + 0.40082c_8 - 0.08578c_9 - 0.18870c_{10} - 0.19267c_{11}$$

2.10.3 การถดถอยทีละขั้น

เป็นวิธีคัดเลือกตัวแปรอิสระที่รวมการเลือกแบบไปข้างหน้ากับการกำจัดแบบถอยหลังเข้าด้วยกัน

จากการใช้โปรแกรม R ทำวิธี Stepwise regression ในการคัดเลือกตัวแบบที่เหมาะสมได้ผลดังรูป ถ. (ภาคผนวก ค.) ได้ตัวแบบการถดถอยที่ประกอบด้วยตัวแปรอิสระจำนวน 5 ตัวแปร ได้แก่ BEDS, BATH, PROPERTYSQFT, TYPE และ COUNTY และได้ค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ดังรูป ท. (ภาคผนวก ค.) ดังนั้นสมการถดถอยที่ประมาณได้ คือ

$$\widehat{\log_{10} y} = 5.47133 - 0.00749x_1 + 0.06769x_2 + 0.00003x_3 + 0.28296d_4 + 0.38534d_5 + 0.38068d_6 + 0.53909d_7 + 0.40082c_8 - 0.08578c_9 - 0.18870c_{10} - 0.19267c_{11}$$

บทที่ 3
สรุปและอภิปรายผลการศึกษา

3.1 สรุปและอภิปรายผล

การศึกษาปัจจัยที่มีอิทธิพลต่อราคาบ้านในเมืองนิวยอร์ก โดยการศึกษาข้อมูลจากเว็บไซต์ Kaggle โดย NIDULA ELGIRIYEWITHANA เป็นผู้รวบรวม มีวัตถุประสงค์เพื่อศึกษาถึงความสัมพันธ์ระหว่างจำนวนห้องนอน จำนวนห้องน้ำ พื้นที่บ้านรวมที่ดิน ประเภทของบ้าน และเขตที่ตั้งของบ้านต่อราคาบ้านในเมืองนิวยอร์ก

การศึกษานี้ใช้ค่าสถิติเบื้องต้น ได้แก่ ค่าเฉลี่ย (Mean) ค่ากลาง (Median) ค่าความแปรปรวน (Variance) ค่ามากสุด (Max) ค่าน้อยสุด (Min) ค่าเบี่ยงเบนมาตรฐาน (Standard deviation) และพิสัย (Range) ในการศึกษาข้อมูลก่อนที่จะทำการวิเคราะห์การถดถอยเชิงเส้น (Multiple Regression Analysis) และจากการศึกษาข้อมูลเบื้องต้น สามารถกำหนดตัวแบบการถดถอยเชิงเส้นพหุคูณ ได้ดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 d_{i4} + \beta_5 d_{i5} + \beta_6 d_{i6} + \beta_7 d_{i7} + \beta_8 c_{i8} + \beta_9 c_{i9} + \beta_{10} c_{i10} + \beta_{11} c_{i11} + \varepsilon_i$$

และได้นำข้อมูลมาประมาณค่าสัมประสิทธิ์การถดถอย ($\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}$และ β_{11}) แล้วได้สมการถดถอยที่ประมาณได้ คือ

$$\hat{y} = -1{,}267{,}000 - 120{,}000x_1 + 502{,}200x_2 + 507.10x_3 + 1{,}019{,}000d_4 + 965{,}000d_5 + 121{,}600d_6 + 3{,}281{,}000d_7 + 2{,}641{,}000c_8 - 88{,}700c_9 - 714{,}200c_{10} - 184{,}400c_{11}$$

ทำการทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอย สรุปได้ว่าจำนวนห้องนอน จำนวนห้องน้ำ พื้นที่บ้านรวมที่ดิน ประเภทของบ้านและเขตที่ตั้งของบ้านสามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้

ทำการตรวจสอบข้อสมมติของตัวแบบการถดถอยทั้ง 5 ข้อสมมติ แล้วพบว่าตัวแบบการถดถอยที่กำหนดผิดข้อสมมติ 3 ข้อ คือ ตัวแปรตามและตัวแปรอิสระไม่ได้มีความสัมพันธ์กันในเชิงเส้น ความคลาดเคลื่อนไม่มีการแจกแจงปกติ และความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ จึงทำการแปลงค่าของตัวแปรตาม y ให้เป็น y' โดยที่ $y' = \log_{10} y$ เพื่อแก้ข้อสมมติของตัวแบบการถดถอยให้เป็นจริง แล้วตรวจสอบข้อสมมติของตัวแบบการถดถอยทั้ง 5 ข้อสมมติอีกครั้ง แล้วพบว่ายังไม่สามารถแก้ข้อสมมติเรื่องความคลาดเคลื่อนต้องมีการแจกแจงปกติได้ เพราะมีค่านอกเกณฑ์ที่เป็นค่าที่มีอิทธิพลทำให้ความคลาดเคลื่อนไม่มีการแจกแจงปกติ

สุดท้ายจะสามารถเลือกตัวแปรอิสระและตัวแบบที่เหมาะสมได้โดยการใช้วิธีการเลือกแบบไปข้างหน้า (Forward selection procedure) วิธีการกำจัดแบบถอยหลัง (Backward elimination procedure) และวิธีการ

ถดถอยทีละขั้น (Stepwise regression procedure) ทั้ง 3 วิธีได้ตัวแบบการถดถอยที่ประกอบด้วยตัวแปรอิสระจำนวน 5 ตัวแปร ได้แก่ BEDS, BATH, PROPERTYSQFT, TYPE และ COUNTY และจะได้สมการถดถอยที่ประมาณได้ คือ

$$\widehat{\log_{10} y} = 5.47133 - 0.00749x_1 + 0.06769x_2 + 0.00003x_3 + 0.28296d_4 + 0.38534d_5 + 0.38068d_6 + 0.53909d_7 + 0.40082c_8 - 0.08578c_9 - 0.18870c_{10} - 0.19267c_{11}$$

โดยที่สัมประสิทธิ์การถดถอยที่ประมาณได้มีความหมายดังนี้

- $\hat{\beta}_1$ = -0.00749 หมายความว่า หากจำนวนห้องนอนในบ้านแต่ละหลังแตกต่างกัน 1 ห้อง ราคาบ้านจะแตกต่างกัน 0.00749 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย โดยกำหนดให้จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินคงที่ เมื่อบ้านจัดอยู่ในประเภทเดียวกันและตั้งอยู่ในเขตเดียวกัน
- $\hat{\beta}_2$ = 0.06769 หมายความว่า หากจำนวนห้องน้ำในบ้านแต่ละหลังแตกต่างกัน 1 ห้อง ราคาบ้านจะแตกต่างกัน 0.06769 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย โดยกำหนดให้จำนวนห้องนอนและพื้นที่บ้านรวมที่ดินคงที่ เมื่อบ้านจัดอยู่ในประเภทเดียวกันและตั้งอยู่ในเขตเดียวกัน
- $\hat{\beta}_3$ = 0.00003 หมายความว่า หากพื้นที่บ้านรวมที่ดินของบ้านแต่ละหลังแตกต่างกัน 1 ตารางฟุต ราคาบ้านจะแตกต่างกัน 0.00003 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย โดยกำหนดให้จำนวนห้องนอนและจำนวนห้องน้ำคงที่ เมื่อบ้านจัดอยู่ในประเภทเดียวกันและตั้งอยู่ในเขตเดียวกัน
- $\hat{\beta}_4$ = 0.28296 หมายความว่า ราคาบ้านในเมืองนิวยอร์กประเภท Condo จะแตกต่างจากราคาบ้านประเภท Co-op อยู่ 0.28296 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย เมื่อบ้านแต่ละประเภทตั้งอยู่ในเขตเดียวกัน มีจำนวนห้องนอน จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินเท่ากัน
- $\hat{\beta}_5$ = 0.38534 หมายความว่า ราคาบ้านในเมืองนิวยอร์กประเภท House จะแตกต่างจากราคาบ้านประเภท Co-op อยู่ 0.38534 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย เมื่อบ้านแต่ละประเภทตั้งอยู่ในเขตเดียวกัน มีจำนวนห้องนอน จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินเท่ากัน
- $\hat{\beta}_6$ = 0.38068 หมายความว่า ราคาบ้านในเมืองนิวยอร์กประเภท Multi – family home จะแตกต่างจากราคาบ้านประเภท Co-op อยู่ 0.38068 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย เมื่อบ้านแต่ละประเภทตั้งอยู่ในเขตเดียวกัน มีจำนวนห้องนอน จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินเท่ากัน

- $\hat{\beta}_7$ = 0.53909 หมายความว่า ราคาบ้านในเมืองนิวยอร์กประเภท Townhouse จะแตกต่างจากราคาบ้านประเภท Co-op อยู่ 0.53909 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย เมื่อบ้านแต่ละประเภทตั้งอยู่ในเขตเดียวกัน มีจำนวนห้องนอน จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินเท่ากัน
- $\hat{\beta}_8$ = 0.40082 หมายความว่า ราคาบ้านในเมืองนิวยอร์กที่ตั้งอยู่ในเขต Manhattan จะแตกต่างจากราคาบ้านในเขต Brooklyn อยู่ 0.40082 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย เมื่อบ้านในแต่ละเขตเป็นประเภทเดียวกัน มีจำนวนห้องนอน จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินเท่ากัน
- $\hat{\beta}_9$ = -0.08578 หมายความว่า ราคาบ้านในเมืองนิวยอร์กที่ตั้งอยู่ในเขต Queens จะแตกต่างจากราคาบ้านในเขต Brooklyn อยู่ 0.08578 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย เมื่อบ้านในแต่ละเขตเป็นประเภทเดียวกัน มีจำนวนห้องนอน จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินเท่ากัน
- $\hat{\beta}_{10}$ = -0.18870 หมายความว่า ราคาบ้านในเมืองนิวยอร์กที่ตั้งอยู่ในเขต Staten Island จะแตกต่างจากราคาบ้านในเขต Brooklyn อยู่ 0.18870 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย เมื่อบ้านในแต่ละเขตเป็นประเภทเดียวกัน มีจำนวนห้องนอน จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินเท่ากัน
- $\hat{\beta}_{11}$ = -0.19267 หมายความว่า ราคาบ้านในเมืองนิวยอร์กที่ตั้งอยู่ในเขต The Bronx จะแตกต่างจากราคาบ้านในเขต Brooklyn อยู่ 0.19267 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย เมื่อบ้านในแต่ละเขตเป็นประเภทเดียวกัน มีจำนวนห้องนอน จำนวนห้องน้ำและพื้นที่บ้านรวมที่ดินเท่ากัน
- ส่วน $\hat{\beta}_0$ = 5.47133 นั้นไม่สามารถแยกอธิบายได้ เนื่องจากค่า $x_1 = 0, x_3 = 0$ ไม่อยู่ในช่วงของข้อมูลที่ใช้ในการประมาณสมการถดถอย

ซึ่งสมการถดถอยที่ประมาณได้นี้มีค่าสัมประสิทธิ์การกำหนด (R^2) เท่ากับ 0.6283 หมายความว่า จำนวนห้องน้ำ จำนวนห้องนอน พื้นที่บ้านรวมที่ดิน ประเภทของบ้าน และเขตที่ตั้งของบ้านสามารถอธิบายความแปรผันของราคาบ้านในเมืองนิวยอร์กได้ร้อยละ 62.83 ดังรูป ฐ. (ภาคผนวก ค.)

และหากนำสมการนี้ไปประมาณค่าของราคาบ้านในเมืองนิวยอร์ก เมื่อกำหนดจำนวนห้องน้ำ จำนวนห้องนอน พื้นที่บ้านรวมที่ดิน ประเภทของบ้าน และเขตที่ตั้งของบ้าน จะเกิดความคลาดเคลื่อน (S) 0.272 ลอการิทึมฐานสิบของดอลลาร์สหรัฐโดยเฉลี่ย ดังรูป ฐ. (ภาคผนวก ค.)

สรุปได้ว่าสมการถดถอยที่ประมาณได้มีค่าความคลาดเคลื่อน เท่ากับ 0.272 ซึ่งถือว่ามีความคลาดเคลื่อนน้อยเมื่อเทียบกับจำนวนข้อมูลทั้งหมด แต่เนื่องจากค่าสัมประสิทธิ์การกำหนด เท่ากับ 0.6283 ซึ่งมีค่าไม่เข้าใกล้ 1 แสดงว่าตัวแปรอิสระทั้งหมดสามารถอธิบายความแปรผันของราคาบ้านได้น้อย ดังนั้นสมการถดถอยที่ประมาณได้อาจจะยังไม่ใช่สมการที่ดีที่สุดในการวิเคราะห์ข้อมูลชุดนี้

3.2 ข้อจำกัดในการทำรายงาน

รายงานนี้ใช้ข้อมูลจากเว็บไซต์ Kaggle ซึ่งเป็นข้อมูลที่มีผู้เก็บรวบรวมข้อมูลไว้แล้ว เมื่อข้อสมมติของตัวแบบการถดถอยไม่เป็นไปตามที่กำหนดหรือมีค่านอกเกณฑ์เกิดขึ้นจึงยากที่จะแก้ไข เนื่องจากไม่ทราบว่าการผิดข้อสมมติหรือค่านอกเกณฑ์เกิดขึ้นจากสาเหตุใด ซึ่งอาจจะเกิดจากความผิดพลาดในการบันทึก การวัดค่าผิด เครื่องมือวัดมีสภาพชำรุด หรือตัวแบบการถดถอยขาดตัวแปรอิสระที่สำคัญไป

3.3 ข้อเสนอแนะ

(1) ในการแก้ข้มสมมติความคลาดเคลื่อนไม่มีการแจกแจงปกติที่เกิดจากค่านอกเกณฑ์ของข้อมูล อาจจะเก็บข้อมูลตัวแปรอิสระที่น่าจะมีอิทธิพลต่อราคาบ้านเพิ่มเติม หรืออาจจะเปลี่ยนไปใช้ non-linear regression

(2) อาจจะจัดกลุ่มข้อมูลของเขตที่ตั้งของบ้านเป็น 2 กลุ่ม เช่น กลุ่มรายได้สูงกับกลุ่มรายได้ปานกลาง เพื่อให้เข้าใจความสัมพันธ์ของตัวแปรตามและตัวแปรอิสระมากขึ้น ตีความได้ง่ายขึ้น ลดความซับซ้อนตัวแบบ และมีความแม่นยำในการประมาณราคาบ้านมากขึ้น

ภาคผนวก

ภาคผนวก ก.

คำสั่งวิเคราะห์ ในโปรแกรม R

1. ศึกษาข้อมูลเชิงสถิติเบื้องต้น

attach(DB_House)

(1) ตัวแปรอิสระ PRICE

mean(PRICE)

median(PRICE)

var(PRICE)

sd(PRICE)

QR(PRICE)

min(PRICE)

max(PRICE)

range(PRICE)

(2) ตัวแปรอิสระ BEDS

mean(BEDS)

median(BEDS)

var(BEDS)

sd(BEDS)

IQR(BEDS)

min(BEDS)

max(BEDS)

range(BEDS)

(3) ตัวแปรอิสระ BATH

mean(BATH)

median(BATH)

var(BATH)

sd(BATH)

IQR(BATH)

min(BATH)

max(BATH)

range(BATH)

(4) ตัวแปรอิสระ PROPERTYSQFT

```
mean(PROPERTYSQFT)
median(PROPERTYSQFT)
var(PROPERTYSQFT)
sd(PROPERTYSQFT)
IQR(PROPERTYSQFT)
min(PROPERTYSQFT)
max(PROPERTYSQFT)
range(PROPERTYSQFT)
```

(5) ตัวแปรอิสระ TYPE

```
table(TYPE)
table(TYPE)/nrow(DB_House)
```

(6) ตัวแปรอิสระ COUNTY

```
table(COUNTY)
table(COUNTY)/nrow(DB_House)*100
```

(7) สร้างแผนภาพดูการกระจายข้อมูลของตัวแปรอิสระ

```
hist(PRICE)
hist(BEDS)
hist(BATH)
hist(PROPERTYSQFT)
barplot(table(COUNTY))
barplot(table(TYPE))
```

(8) สร้างแผนภาพดูความสัมพันธ์ Y กับ X เชิงปริมาณ

```
plot(BEDS,PRICE,xlab = "Beds",ylab = "Price",col="blue",pch=20)
plot(BATH,PRICE,xlab = "Bath",ylab = "Price",col="blue",pch=20)
plot(PROPERTYSQFT,PRICE,xlab = "Propertysqft",ylab = "Price",col="blue",pch=20)
```

2. Original Model

2.1 สร้าง model

```
model = lm(PRICE ~ BEDS+BATH+PROPERTYSQFT+TYPE+COUNTY, data=DB_House)
summary(model)
anova(model)
```

2.2 ตรวจสอบข้อสมมติ

(1) การตรวจสอบความสัมพันธ์เชิงเส้นของตัวแบบ และความคงที่ของความแปรปรวนของความคลาดเคลื่อน

```
y.hat = fitted.values(model)
res = resid(model)
plot(y.hat, res, ylab="Residuals", xlab="Fitted Values",col="blue",pch=20)
```

(2) การตรวจสอบการแจกแจงปกติของความคลาดเคลื่อน

```
library(nortest)
ad.test(res)
qqnorm(res,ylab="Residuals",xlab="Expected value",col="blue",pch=20 )
```

(3) การตรวจสอบความเป็นอิสระเชิงเส้นกันของความคลาดเคลื่อน

```
plot(i, res, ylab="Residual", xlab="Time Order",col="blue",pch=20)
library(lmtest)
dwtest(model)
```

(4) การตรวจสอบความเป็นอิสระกันของตัวแปรอิสระในตัวแบบ

```
library(car)
model = lm(PRICE ~ BEDS+BATH+PROPERTYSQFT+TYPE+COUNTY, data=DB_House)
vif(model)
```

3. New Model แปลงค่าตัวแปรตาม ($\log_{10} y$)

3.1 สร้าง model

```
new.model = lm(log10(PRICE) ~ BEDS+BATH+PROPERTYSQFT+TYPE+COUNTY, data=DB_House)
summary(new.model)
anova(new.model)
```

3.2 ตรวจสอบข้อสมมติ

(1) การตรวจสอบความสัมพันธ์เชิงเส้นของตัวแบบ และความคงที่ของความแปรปรวนของความคลาดเคลื่อน

```
new.y.hat = fitted.values(new.model)
new.res = resid(new.model)
plot(new.y.hat, new.res, ylab="Residuals", xlab="Fitted Values",col="blue",pch=20)
```

(2) การตรวจสอบการแจกแจงปกติของความคลาดเคลื่อน

```
library(nortest)
ad.test(new.res)
qqnorm(new.res,ylab="Residuals",xlab="Expected value",col="blue",pch=20 )
```

(3) การตรวจสอบความเป็นอิสระเชิงเส้นกันของความคลาดเคลื่อน

```
plot(i, new.res, ylab="Residual", xlab="Time Order",col="blue",pch=20)
library(lmtest)
dwtest(new.model)
```

(4) การตรวจสอบความเป็นอิสระกันของตัวแปรอิสระในตัวแบบ

```
library(car)
vif(new.model)
```

3.3 ตรวจสอบค่านอกเกณฑ์

```
new.rstandard = rstandard(new.model)
write.csv(new.rstandard, file = "stdres")
plot(new.y.hat, new.rstandard, ylab="Standardized Residual", xlab="Fitted
Values",col="blue",pch=20)
```

3.4 ตรวจสอบค่าที่มีอิทธิพล

```
hatvalues = hatvalues(new.model)
write.csv(hatvalues, file = "hatvalues")
cooks = cooks.distance(new.model)
write.csv(cooks, file = "cooks")
diffits = dffits(new.model)
write.csv(diffits, file = "diffits")
```

4. เลือกตัวแบบที่เหมาะสม

(1) แบบไปข้างหน้า

```
model.start = lm(log10(PRICE)~1, data=DB_House)
model.forw = step(model.start,scope = log10(PRICE) ~
BEDS+BATH+PROPERTYSQFT+TYPE+COUNTY, data=DB_House, direction = "forward")
coef(model.forw)
```

(2) แบบถอยหลัง

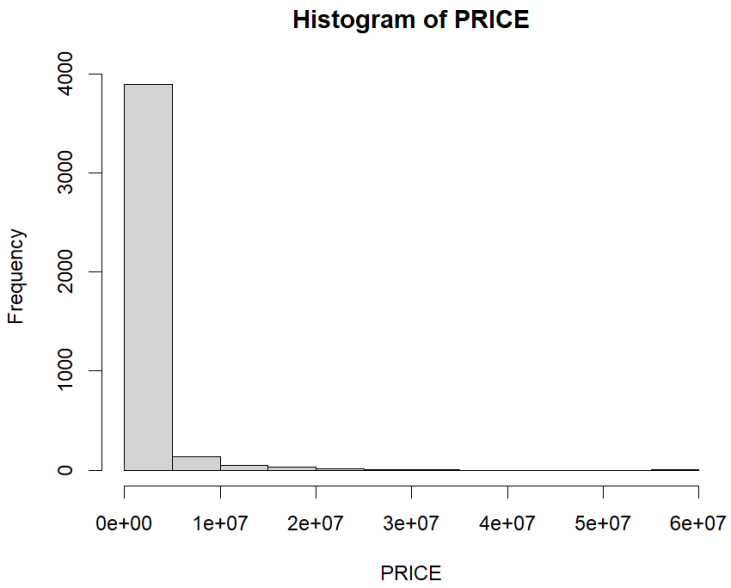
```
model.back = step(new.model, direction = "backward")
coef(model.back)
```

(3) stepwise

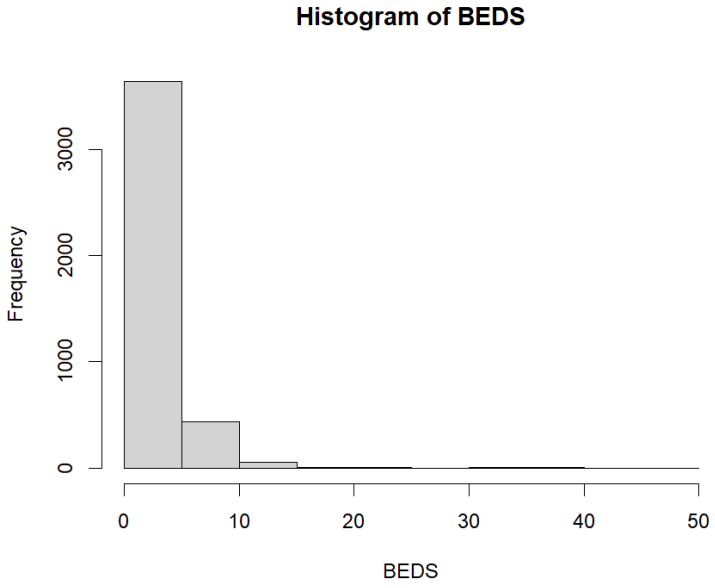
```
model.step = step(model.start,scope = log10(PRICE) ~
BEDS+BATH+PROPERTYSQFT+TYPE+COUNTY, data=DB_House ,direction = "both")
coef(model.step)
```

ภาคผนวก ข.

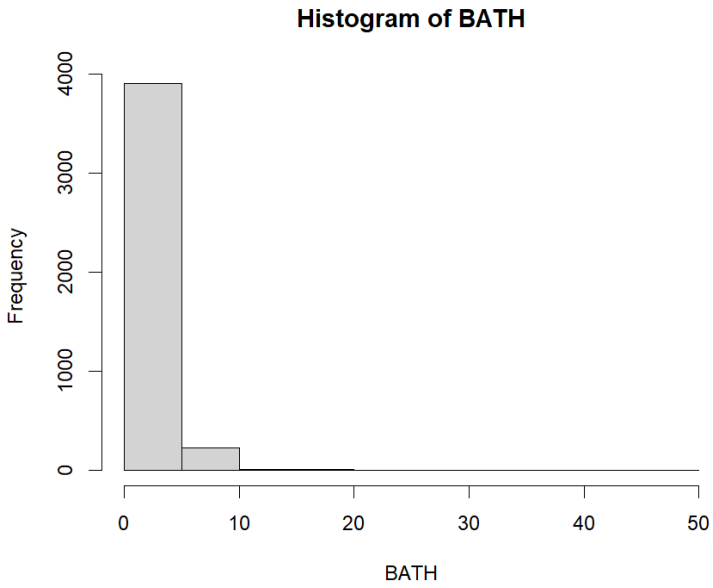
แผนภาพแสดงข้อมูล



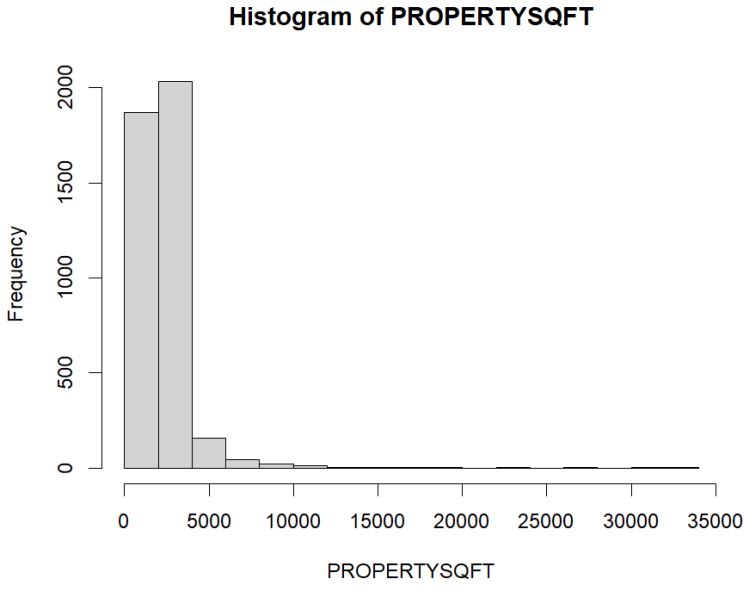
รูป ก. แผนภาพฮิสโทแกรมของราคาบ้าน



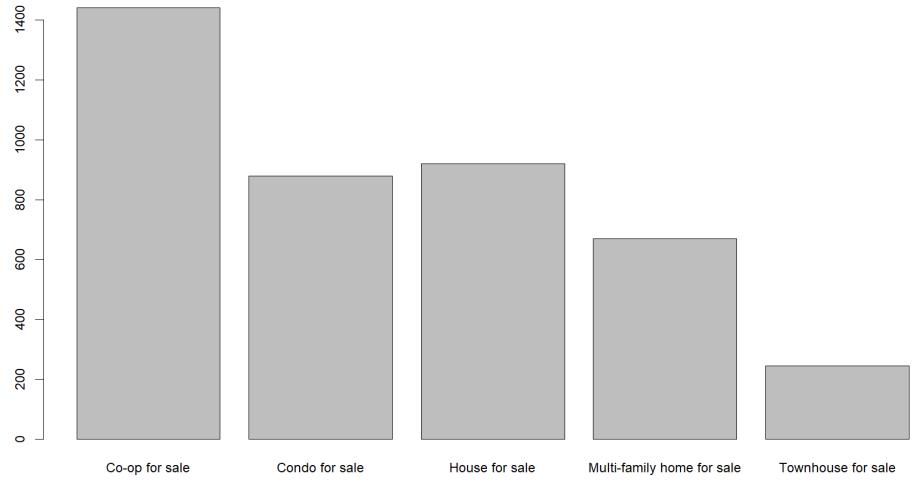
รูป ข. แผนภาพฮิสโทแกรมของจำนวนห้องนอน



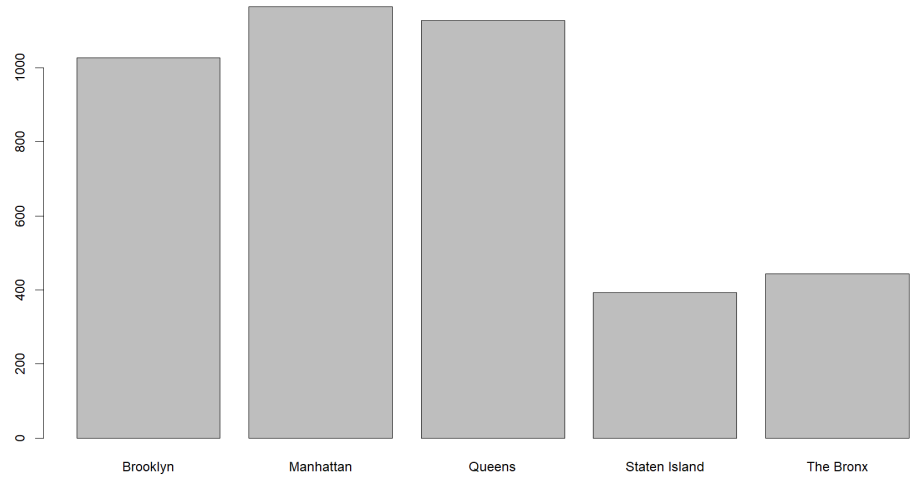
รูป ค. แผนภาพฮิสโทแกรมของจำนวนห้องน้ำ



รูป ง. แผนภาพฮิสโทแกรมของพื้นที่บ้านรวมที่ดิน



รูป จ. แผนภูมิแท่งแสดงประเภทของบ้าน



รูป ฉ. แผนภูมิแท่งแสดงเขตที่ตั้งของบ้านในเมืองนิวยอร์ก

ภาคผนวก ค.

ผลการวิเคราะห์จากโปรแกรม R

```
> summary(model)

Call:
lm(formula = PRICE ~ BEDS + BATH + PROPERTYSQFT + TYPE + COUNTY,
    data = DB_House)

Residuals:
      Min        1Q    Median        3Q       Max
-21816935  -1140246    -38258    536592  54459946

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.267e+06  1.440e+05  -8.797  < 2e-16 ***
BEDS                  -1.200e+05  3.370e+04  -3.560 0.000374 ***
BATH                   5.022e+05  4.304e+04  11.668  < 2e-16 ***
PROPERTYSQFT           5.071e+02  2.398e+01  21.145  < 2e-16 ***
TYPECondo              1.019e+06  1.458e+05   6.992 3.11e-12 ***
TYPEHouse              9.650e+05  1.600e+05   6.031 1.76e-09 ***
TYPEMulti-family home  1.216e+05  1.949e+05   0.624 0.532906
TYPETownhouse          3.281e+06  2.314e+05  14.178  < 2e-16 ***
COUNTYManhattan        2.641e+06  1.518e+05  17.399  < 2e-16 ***
COUNTYQueens          -8.870e+04  1.438e+05  -0.617 0.537308
COUNTYStaten Island   -7.142e+05  1.989e+05  -3.591 0.000333 ***
COUNTYThe Bronx       -1.844e+05  1.868e+05  -0.987 0.323494
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3368000 on 4371 degrees of freedom
Multiple R-squared:  0.3582,    Adjusted R-squared:  0.3566
F-statistic: 221.8 on 11 and 4371 DF,  p-value: < 2.2e-16

> anova(model)
Analysis of Variance Table

Response: PRICE
               Df     Sum Sq    Mean Sq F value    Pr(>F)
BEDS            1 4.2222e+15 4.2222e+15  372.17 < 2.2e-16 ***
BATH            1 6.7637e+15 6.7637e+15  596.19 < 2.2e-16 ***
PROPERTYSQFT    1 6.7839e+15 6.7839e+15  597.96 < 2.2e-16 ***
TYPE            4 4.5499e+15 1.1375e+15  100.26 < 2.2e-16 ***
COUNTY          4 5.3573e+15 1.3393e+15  118.05 < 2.2e-16 ***
Residuals    4371 4.9589e+16 1.1345e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

รูป ช. ผลการวิเคราะห์การถอถอยของข้อมูล

```
        Anderson-Darling normality test

data:  res
A = 506.65, p-value < 2.2e-16
```

รูป ซ. ผลการทดสอบ Anderson-Darling

```
        Durbin-Watson test

data:  model
DW = 2.0462, p-value = 0.9367
alternative hypothesis: true autocorrelation is greater than 0
```

รูป ฌ. ผลการทดสอบ Durbin-Watson

```
> vif(model)
                 GVIF Df GVIF^(1/(2*Df))
BEDS         3.143773  1        1.773069
BATH         2.867375  1        1.693333
PROPERTYSQFT 1.342201  1        1.158534
TYPE         2.173051  4        1.101879
COUNTY       1.462001  4        1.048621
```

รูป ญ. ผลการวิเคราะห์ Variance inflation factor (VIF)

```
        Anderson-Darling normality test

data:  new.res
A = 36.798, p-value < 2.2e-16
```

รูป ฎ. ผลการทดสอบ Anderson-Darling หลังจากแปลงค่าตัวแปรตาม

```
        Durbin-Watson test

data:  new.model
DW = 1.961, p-value = 0.0979
alternative hypothesis: true autocorrelation is greater than 0
```

รูป ฏ. ผลการทดสอบ Durbin-Watson หลังจากแปลงค่าตัวแปรตาม

```
> summary(new.model)

Call:
lm(formula = log10(PRICE) ~ BEDS + BATH + PROPERTYSQFT + TYPE +
    COUNTY, data = DB_House)

Residuals:
     Min       1Q   Median       3Q      Max
-2.55426 -0.14868 -0.00863  0.13319  1.61313

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.471e+00  1.163e-02 470.484  < 2e-16 ***
BEDS                 -7.494e-03  2.722e-03  -2.753  0.00593 **
BATH                  6.769e-02  3.476e-03  19.470  < 2e-16 ***
PROPERTYSQFT          3.331e-05  1.937e-06  17.197  < 2e-16 ***
TYPECondo             2.830e-01  1.178e-02  24.029  < 2e-16 ***
TYPEHouse             3.853e-01  1.292e-02  29.818  < 2e-16 ***
TYPEMulti-family home 3.807e-01  1.574e-02  24.181  < 2e-16 ***
TYPETownhouse         5.391e-01  1.869e-02  28.844  < 2e-16 ***
COUNTYManhattan       4.008e-01  1.226e-02  32.692  < 2e-16 ***
COUNTYQueens         -8.578e-02  1.161e-02  -7.387 1.79e-13 ***
COUNTYStaten Island  -1.887e-01  1.606e-02 -11.748  < 2e-16 ***
COUNTYThe Bronx      -1.927e-01  1.508e-02 -12.773  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.272 on 4371 degrees of freedom
Multiple R-squared:  0.6283,    Adjusted R-squared:  0.6274
F-statistic: 671.7 on 11 and 4371 DF,  p-value: < 2.2e-16

> anova(new.model)
Analysis of Variance Table

Response: log10(PRICE)
               Df Sum Sq Mean Sq F value    Pr(>F)
BEDS            1 145.17 145.167 1961.55 < 2.2e-16 ***
BATH            1 105.17 105.173 1421.14 < 2.2e-16 ***
PROPERTYSQFT    1  37.47  37.473  506.35 < 2.2e-16 ***
TYPE            4  85.67  21.419  289.42 < 2.2e-16 ***
COUNTY          4 173.30  43.324  585.41 < 2.2e-16 ***
Residuals    4371 323.48   0.074
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

รูป ฐ. ผลการวิเคราะห์การถดถอยของข้อมูล หลังจากแปลงค่าตัวแปรตาม

```
> vif(new.model)
                    GVIF Df GVIF^(1/(2*Df))
BEDS           3.143773  1        1.773069
BATH           2.867375  1        1.693333
PROPERTYSQFT   1.342201  1        1.158534
TYPE           2.173051  4        1.101879
COUNTY         1.462001  4        1.048621
```

รูป ฑ. ผลการวิเคราะห์ Variance inflation factor (VIF) หลังจากแปลงค่าตัวแปรตาม

```
> model.start = lm(log10(PRICE)~1, data=DB_House)
> model.forw = step(model.start,scope = log10(PRICE) ~ BEDS+BATH+PROPERTYSQFT+TYPE+COUNTY,
data=DB_House, direction = "forward" )
Start:  AIC=-7083.97
log10(PRICE) ~ 1

               Df Sum of Sq    RSS     AIC
+ BATH          1    250.13 620.13 -8567.2
+ TYPE          4    195.68 674.58 -8192.3
+ COUNTY        4    184.85 685.41 -8122.6
+ PROPERTYSQFT  1    166.34 703.92 -8011.8
+ BEDS          1    145.17 725.10 -7881.8
<none>                      870.26 -7084.0

Step:  AIC=-8567.19
log10(PRICE) ~ BATH

               Df Sum of Sq    RSS     AIC
+ COUNTY        4   161.644 458.49 -9882.8
+ TYPE          4    88.202 531.93 -9231.6
+ PROPERTYSQFT  1    36.784 583.35 -8833.2
<none>                      620.13 -8567.2
+ BEDS          1     0.212 619.92 -8566.7

Step:  AIC=-9882.84
log10(PRICE) ~ BATH + COUNTY

               Df Sum of Sq    RSS      AIC
+ TYPE          4   112.891 345.60 -11113.7
+ PROPERTYSQFT  1    28.693 429.80 -10164.1
+ BEDS          1     3.090 455.40  -9910.5
<none>                      458.49  -9882.8

Step:  AIC=-11113.73
log10(PRICE) ~ BATH + COUNTY + TYPE

               Df Sum of Sq    RSS    AIC
+ PROPERTYSQFT  1   21.5590 324.04 -11394
+ BEDS          1    0.2342 345.37 -11115
<none>                      345.60 -11114

Step:  AIC=-11394.04
log10(PRICE) ~ BATH + COUNTY + TYPE + PROPERTYSQFT

       Df Sum of Sq    RSS    AIC
+ BEDS  1   0.56091 323.48 -11400
<none>              324.04 -11394

Step:  AIC=-11399.64
log10(PRICE) ~ BATH + COUNTY + TYPE + PROPERTYSQFT + BEDS
```

รูป ฒ. Forward selection ด้วยเกณฑ์ AIC

```
> coef(model.forw)
          (Intercept)                 BATH       COUNTYManhattan          COUNTYQueens
         5.471325e+00         6.768615e-02         4.008195e-01        -8.577820e-02
  COUNTYStaten Island       COUNTYThe Bronx          TYPECondo          TYPEHouse
        -1.887011e-01        -1.926702e-01         2.829609e-01         3.853394e-01
TYPEMulti-family home       TYPETownhouse       PROPERTYSQFT                 BEDS
         3.806828e-01         5.390861e-01         3.331196e-05        -7.493800e-03
```

รูป ณ. สัมประสิทธิ์การถดถอยที่ประมาณได้จาก Forward selection

```
> model.back = step(new.model, direction = "backward")
Start:  AIC=-11399.64
log10(PRICE) ~ BEDS + BATH + PROPERTYSQFT + TYPE + COUNTY

               Df Sum of Sq    RSS      AIC
<none>                      323.48 -11399.6
- BEDS          1     0.561 324.04 -11394.0
- PROPERTYSQFT  1    21.886 345.37 -11114.7
- BATH          1    28.055 351.54 -11037.1
- TYPE          4   104.746 428.23 -10178.1
- COUNTY        4   173.296 496.78  -9527.3
```

รูป ด. Backward elimination ด้วยเกณฑ์ AIC

```
> coef(model.back)
          (Intercept)                 BEDS                 BATH       PROPERTYSQFT
         5.471325e+00        -7.493800e-03         6.768615e-02         3.331196e-05
             TYPECondo            TYPEHouse TYPEMulti-family home       TYPETownhouse
         2.829609e-01         3.853394e-01         3.806828e-01         5.390861e-01
       COUNTYManhattan          COUNTYQueens   COUNTYStaten Island       COUNTYThe Bronx
         4.008195e-01        -8.577820e-02        -1.887011e-01        -1.926702e-01
```

รูป ต. สัมประสิทธิ์การถดถอยที่ประมาณได้จาก Backward elimination

```
> model.step = step(model.start,scope = log10(PRICE) ~ BEDS+BATH+PROPERTYSQFT+TYPE+COUNTY,
data=DB_House ,direction = "both")
Start:  AIC=-7083.97
log10(PRICE) ~ 1

               Df Sum of Sq    RSS     AIC
+ BATH          1    250.13 620.13 -8567.2
+ TYPE          4    195.68 674.58 -8192.3
+ COUNTY        4    184.85 685.41 -8122.6
+ PROPERTYSQFT  1    166.34 703.92 -8011.8
+ BEDS          1    145.17 725.10 -7881.8
<none>                      870.26 -7084.0

Step:  AIC=-8567.19
log10(PRICE) ~ BATH

               Df Sum of Sq    RSS     AIC
+ COUNTY        4   161.644 458.49 -9882.8
+ TYPE          4    88.202 531.93 -9231.6
+ PROPERTYSQFT  1    36.784 583.35 -8833.2
<none>                      620.13 -8567.2
+ BEDS          1     0.212 619.92 -8566.7
- BATH          1   250.128 870.26 -7084.0

Step:  AIC=-9882.84
log10(PRICE) ~ BATH + COUNTY

               Df Sum of Sq    RSS      AIC
+ TYPE          4   112.891 345.60 -11113.7
+ PROPERTYSQFT  1    28.693 429.80 -10164.1
+ BEDS          1     3.090 455.40  -9910.5
<none>                      458.49  -9882.8
- COUNTY        4   161.644 620.13  -8567.2
- BATH          1   226.916 685.41  -8122.6

Step:  AIC=-11113.73
log10(PRICE) ~ BATH + COUNTY + TYPE

               Df Sum of Sq    RSS      AIC
+ PROPERTYSQFT  1    21.559 324.04 -11394.0
+ BEDS          1     0.234 345.37 -11114.7
<none>                      345.60 -11113.7
- BATH          1    80.944 426.54 -10193.4
- TYPE          4   112.891 458.49  -9882.8
- COUNTY        4   186.332 531.93  -9231.6

Step:  AIC=-11394.04
log10(PRICE) ~ BATH + COUNTY + TYPE + PROPERTYSQFT

               Df Sum of Sq    RSS      AIC
+ BEDS          1     0.561 323.48 -11399.6
<none>                      324.04 -11394.0
- PROPERTYSQFT  1    21.559 345.60 -11113.7
- BATH          1    41.945 365.99 -10862.5
- TYPE          4   105.756 429.80 -10164.1
- COUNTY        4   173.515 497.56  -9522.4

Step:  AIC=-11399.64
log10(PRICE) ~ BATH + COUNTY + TYPE + PROPERTYSQFT + BEDS

               Df Sum of Sq    RSS      AIC
<none>                      323.48 -11399.6
- BEDS          1     0.561 324.04 -11394.0
- PROPERTYSQFT  1    21.886 345.37 -11114.7
- BATH          1    28.055 351.54 -11037.1
- TYPE          4   104.746 428.23 -10178.1
- COUNTY        4   173.296 496.78  -9527.3
```

รูป ถ. Stepwise regression ด้วยเกณฑ์ AIC

```
> coef(model.step)
          (Intercept)                  BATH       COUNTYManhattan          COUNTYQueens
         5.471325e+00          6.768615e-02          4.008195e-01         -8.577820e-02
  COUNTYStaten Island       COUNTYThe Bronx             TYPECondo             TYPEHouse
        -1.887011e-01         -1.926702e-01          2.829609e-01          3.853394e-01
TYPEMulti-family home         TYPETownhouse          PROPERTYSQFT                  BEDS
         3.806828e-01          5.390861e-01          3.331196e-05         -7.493800e-03
```

รูป ท. สัมประสิทธิ์การถดถอยที่ประมาณได้จาก Stepwise regression