Assignment 2

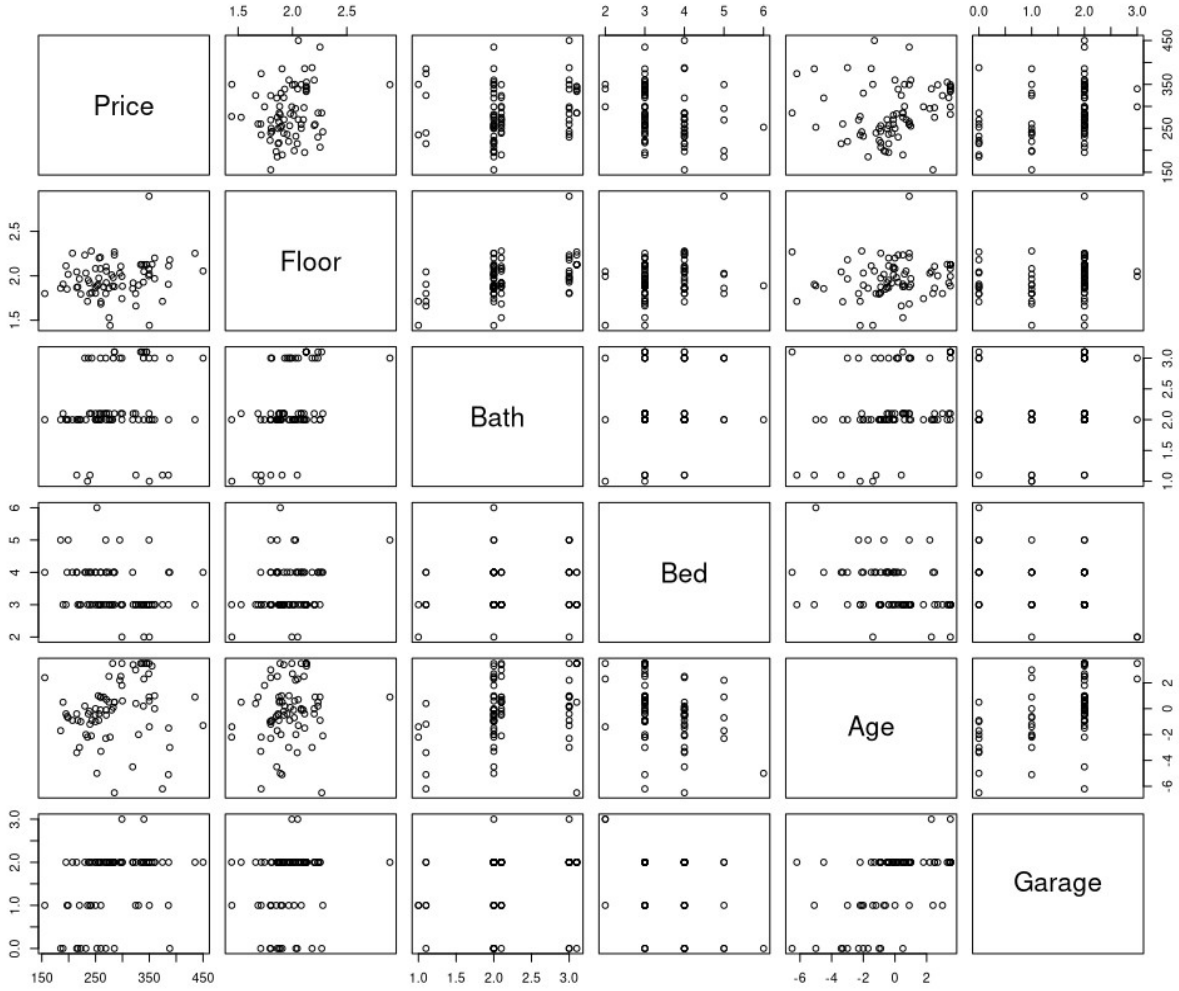
การวิเคราะห์การถดถอยพหุคูณ

ข้อ 1. จงสร้างเมทริกซ์แผนภาพการกระจาย (matrix of scatter plots) ระหว่างตัวแปร Price, Floor, Bath, Bed, Age และ Garage

คำสั่ง R

```
pairs(Homes[,c('Price', 'Floor', 'Bath', 'Bed', 'Age', 'Garage')])
```

จากเมทริกซ์แผนภาพการกระจาย อธิบายความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ได้ดังนี้



ความสัมพันธ์ระหว่างตัวแปรตอบสนองกับตัวแปรอธิบาย (Price vs. Floor, Bath, Bed, Age และ Garage)

Price vs. Floor : Price มีความสัมพันธ์เชิงเส้นเชิงบวกกับ Floor

Price vs. Bath : Price ไม่มีความสัมพันธ์เชิงเส้นกับ Bath

Price vs. Bed : Price มีความสัมพันธ์เชิงเส้นเชิงลบกับ Bed

Price vs. Age : Price มีความสัมพันธ์เชิงเส้นเชิงบวกกับ Age

Price vs. Garage : Price มีความสัมพันธ์เชิงเส้นเชิงบวกกับ Garage

ความสัมพันธ์ระหว่างตัวแปรทำนายต่าง ๆ (Floor, Bath, Bed, Age และ Garage)

Floor กับ Bath มีความสัมพันธ์เชิงเส้นเชิงบวกกัน

Bed กับ Age มีความสัมพันธ์เชิงเส้นเชิงลบกัน

Age กับ Garage มีความสัมพันธ์เชิงเส้นเชิงบวกกัน

ข้อ 2. จงวิเคราะห์ข้อมูลด้วยวิธีการถดถอยพหูคูณข้อมูล โดยใช้ชุดข้อมูล Homes เพื่อหาสมการถดถอยที่เหมาะสมในการทำนายราคาบ้าน (Price) ด้วยตัวแปรทำนายต่าง ๆ ด้วยตัวแบบดังนี้

$$Price = \beta_0 + \beta_1 Floor + \beta_2 Lot + \beta_3 Bath + \beta_4 Bed + \beta_5 Age + \beta_6 Garage + \beta_7 Active + \beta_8 D_{Ed} + \beta_9 D_{Ha} + \beta_{10} D_{Cr} + \beta_{11} D_{Pa} + \varepsilon_i$$

2.1 จงหาสมการถดถอยของตัวอย่างสำหรับการทำนายค่าเฉลี่ยของราคาบ้าน (Price) ตามตัวแบบการถดถอยพหุคูณที่กำหนดข้างต้น

คำสั่ง R

```
model <- lm(Price ~ Floor + Lot + Bath + Bed + Age + Garage + Active + DEd + DHa + DAd + DCr + DPa,  data = Homes)
```

จากผลลัพธ์ของคำสั่ง R ข้างต้น

สมการถดถอยพหุคูณของตัวอย่าง คือ

$$Price = 95.607 + 69.785\,Floor + 10.749\,Lot + 4.616\,Bath - 12.460\,Bed + 1.626\,Age + 10.052\,Garage + 30.876\,Active + 79.730\,D_{Ed} + 46.322\,D_{Ha} - 7.465\,D_{Ad} - 2.658\,D_{Cr} - 19.369\,D_{Pa} + \varepsilon_i$$

R^2 มีค่าเท่ากับ 0.5305 หมายถึง สมการเชิงเส้นนี้สามารถอธิบายความแปรปรวนของ Price ได้ 0.5305

ความคลาดเคลื่อนมาตรฐานของส่วนเหลือ (Residual standard error) มีค่าเท่ากับ 45.11

ค่าสถิติทดสอบ F มีค่าเท่ากับ 5.932 ค่า p-value เท่ากับ 8.902e-07 ซึ่งสรุปได้ว่า

มีค่า β ตัวใดตัวหนึ่ง ไม่เท่ากับ 0

2.2 จงเขียนอธิบายผลการทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอยของสมการถดถอยในข้อ 2.1

สมมติฐานการทดสอบ คือ $H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$

จงเติมตัวเลขลงในตาราง

ตัวแปร	ค่าประมาณสัมประสิทธิ์การถดถอย $\hat{\beta}_j$	ความคลาดเคลื่อนมาตรฐานของ $\hat{\beta}_j$ $SE(\hat{\beta}_j)$	สถิติทดสอบที t-statistic	p-value
Intercept	95.607	58.390	1.637	0.10653
Floor	69.785	30.934	2.256	0.02755
Lot	10.749	3.667	2.932	0.00469
Bath	4.616	11.748	0.393	0.69570
Bed	-12.460	9.120	-1.366	0.17674
Age	1.626	3.324	0.489	0.62640
Garage	10.052	9.261	1.085	0.28184
Active	30.876	12.902	2.393	0.01970
DEd	79.730	17.626	4.524	2.75e-05
DHa	46.322	16.291	2.843	0.00601
DAd	-7.465	29.050	-0.257	0.79803
DCr	-2.658	22.925	-0.116	0.90808
DPa	-19.369	15.646	-1.238	0.22033

จากผลการทดสอบสมมติฐานข้างต้น จงสรุปว่า ตัวแปรทำนายตัวใดบ้างที่มีผลต่อราคาขายของบ้าน อย่างมีนัยสำคัญทางสถิติ และจงอธิบายค่าสัมประสิทธิ์ของถดถอยตัวอย่างของตัวแปรทำนายดังกล่าว

ตัวแปร DEd มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.001

ตัวแปร Lot และ DHa มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.01

ตัวแปร Floor และ Active มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05

ค่าสัมประสิทธิ์การถดถอยของ DEd มีความหมายว่า บ้านที่อยู่ใกล้โรงเรียน Edison จะมีราคามากกว่า บ้านที่อยู่ใกล้โรงเรียน Redwood อยู่ 79.73 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ Lot มีความหมายว่า บ้านที่ถูกจัดอยู่ในกลุ่มที่ใหญ่ขึ้น 1 จะมีราคามากขึ้น 10.749 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ DHa มีความหมายว่า บ้านที่อยู่ใกล้โรงเรียน Harris จะมีราคามากกว่า บ้านที่อยู่ใกล้โรงเรียน Redwood อยู่ 46.322 หน่วย

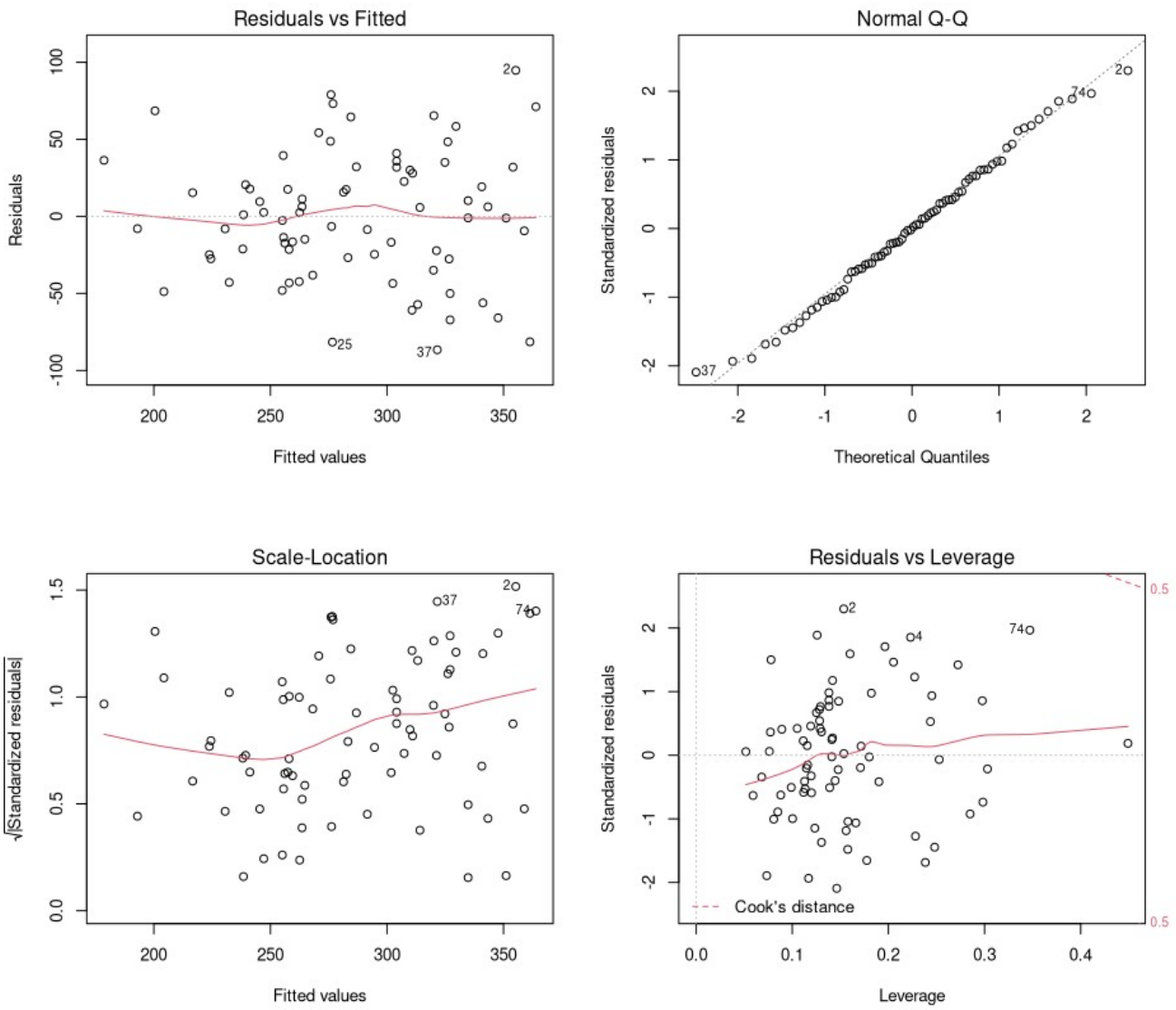
ค่าสัมประสิทธิ์การถดถอยของ Floor มีความหมายว่า บ้านที่มีขนาดมากขึ้น 1000 ตารางฟุต จะมีราคามากขึ้น 69.785 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ Active มีความหมายว่า บ้านที่อยู่ในรายการขายจะมีราคามากกว่าบ้านที่ขายแล้วอยู่ 30.876 หน่วย

2.3 จงตรวจสอบข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยด้วยแผนภาพของส่วนเหลือ (residual plot) และอธิบายลักษณะของแผนภาพว่ามีความสอดคล้องกับข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอย หรือไม่อย่างไร

คำสั่ง R

```
par(mfrow=c(2,2))
plot(model)
```



จากแผนภาพ

Residual vs fitted plot : สอดคล้อง เพราะ ค่าเฉลี่ยของความคลาดเคลื่อนเท่ากับ 0 และความคลาดเคลื่อนมีการกระจายคงที่

Normal Q-Q plot : สอดคล้อง เพราะ การกระจายของความคลาดเคลื่อนเป็นการแจงปกติ

Scale-location plot : ไม่สอดคล้อง เพราะ ค่าความคลาดเคลื่อนไม่คงที่

Residual vs leverage plot : สอดคล้อง เพราะ ไม่มีค่านอกกลุ่ม

ข้อ 3. จงวิเคราะห์ข้อมูลด้วยวิธีการถดถอยพหูคูณข้อมูลโดยใช้ชุดข้อมูล Homes เพื่อหาตัวแบบที่เหมาะสม

สำหรับการทำนายราคาบ้าน (Price) ตามตัวแบบการถดถอย polynomial regression ที่มีเทอมของ $Bath*Bed$ และ Age^2 ดังนี้

$$Price=\beta_0+\beta_1 Floor+\beta_2 Lot+\beta_3 Bath+\beta_4 Bed+\beta_5 Bath*Bed+\beta_6 Age$$
$$+\beta_7 Age^2+\beta_8 Garage+\beta_9 Active+\beta_{10} D_{Ed}+\beta_{11} D_{Ha}+\beta_{12} D_{Ad}+\beta_{13} D_{Cr}+\beta_{14} D_{Pa}+\varepsilon_i$$

3.1 จงหาสมการถดถอยของตัวอย่างสำหรับการทำนายราคาบ้าน (Price) ตามตัวแบบการถดถอย polynomial regression ข้างต้น

คำสั่ง R

```
model ← lm(Price ~ Floor + Lot + Bath + Bed + I(Bath*Bed) + Age + I(Age^2) +
Garage + Active + DEd + DHa + DAd + DCr + DPa, data = Homes)
```

จากผลลัพธ์ของคำสั่ง R ข้างต้น

สมการถดถอยพหุคูณของตัวอย่าง คือ

$$Price=337.6528+58.7715\,Floor+10.3614\,Lot-98.7324\,Bath-77.4802\,Bed$$
$$+29.6564\,Bath*Bed+3.7773\,Age+1.8237\,Age^2+10.6769\,Garage+30.3595\,Active$$
$$+59.2151\,D_{Ed}+40.2357\,D_{Ha}-28.8884\,D_{Ad}-8.8829\,D_{Cr}+-13.9343\,D_{Pa}+\varepsilon_i$$

R^2 มีค่าเท่ากับ 0.5989 หมายถึง สมการเชิงเส้นนี้สามารถอธิบายความแปรปรวนของ Price ได้ 0.5989

ความคลาดเคลื่อนมาตรฐานของส่วนเหลือ (Residual standard error) มีค่าเท่ากับ 42.37

ค่าสถิติทดสอบ F มีค่าเท่ากับ 6.506 ค่า p-value เท่ากับ 8.901e-08 ซึ่งสรุปได้ว่า

มีค่า β ตัวใดตัวหนึ่ง ไม่เท่ากับ 0

3.2 จงทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอยของสมการถดถอยในข้อ 3.1

สมมติฐานการทดสอบ คือ $H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$

ค่าสถิติทดสอบทีสำหรับทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอยต่าง ๆ มีดังนี้

ตัวแปร	ค่าประมาณสัมประสิทธิ์การถดถอย $\hat{\beta}_j$	ความคลาดเคลื่อนมาตรฐานของ $\hat{\beta}_j$ $SE(\hat{\beta}_j)$	สถิติทดสอบที t-statistic	p-value
Intercept	337.6528	124.9619	2.702	0.00891
Floor	58.7715	29.2568	2.009	0.04899
Lot	10.3614	3.5731	2.900	0.00518
Bath	-98.7324	47.9506	-2.059	0.04377
Bed	-77.4802	32.3251	-2.397	0.01961
Bath*Bed	29.6564	13.6582	2.171	0.03381
Age	3.7773	3.2371	1.167	0.24781
Age^2	1.8237	0.7571	2.409	0.01904
Garage	10.6769	8.7030	1.227	0.22461
Active	30.3595	12.2685	2.475	0.01614
DEd	59.2151	18.2076	3.252	0.00187
DHa	40.2357	16.3716	2.458	0.01684
DAd	-28.8884	28.2176	-1.024	0.30998
DCr	-8.8829	21.6213	-0.411	0.68263
DPa	-13.9343	16.0736	-0.867	0.38939

จากผลการทดสอบสมมติฐานข้างต้น จงสรุปว่า ตัวแปรทำนายตัวใดบ้างที่มีผลต่อราคาขายของบ้าน อย่างมีนัยสำคัญทางสถิติ

ตัวแปร Intercept , Lot, DEd มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.01

ตัวแปร Floor , Bath , Bed , Bath * Bed , Age^2 , Active , DHa มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05

ค่าสัมประสิทธิ์การถดถอยของ Intercept มีความหมายว่า บ้านทุกบ้านจะมีราคาเพิ่มขึ้น 337.6528 หน่วย จากค่าอื่นๆ

ค่าสัมประสิทธิ์การถดถอยของ Lot มีความหมายว่า บ้านที่ถูกจัดอยู่ในกลุ่มที่ใหญ่ขึ้น 1 กลุ่ม จะมีราคามากขึ้น 10.3614 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ DEd มีความหมายว่า บ้านที่อยู่ใกล้โรงเรียน Edison จะมีราคามากกว่า บ้านที่อยู่ใกล้โรงเรียน Redwood อยู่ 59.2151 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ Floor มีความหมายว่า บ้านที่มีขนาดมากขึ้น 1000 ตารางฟุต จะมีราคามากขึ้น 58.7715 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ Bath มีความหมายว่า บ้านที่มีห้องน้ำมากขึ้น 1 ห้อง จะมีราคาน้อยขึ้น 98.7324 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ Bed มีความหมายว่า บ้านที่มีห้องนอนมากขึ้น 1 ห้อง จะมีราคาน้อยขึ้น 77.4802 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ Bath*Bed มีความหมายว่า บ้านที่มีผลคูณของจำนวนห้องน้ำกับห้องนอนมากขึ้น 1 หน่วย จะมีราคามากขึ้น 29.6564 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ Age^2 มีความหมายว่า บ้านที่มีค่ากำลังสองของปีที่สร้างลบ 1970 หาร 10 มากขึ้น 1 หน่วย จะมีราคามากขึ้น 1.8237 หน่วย

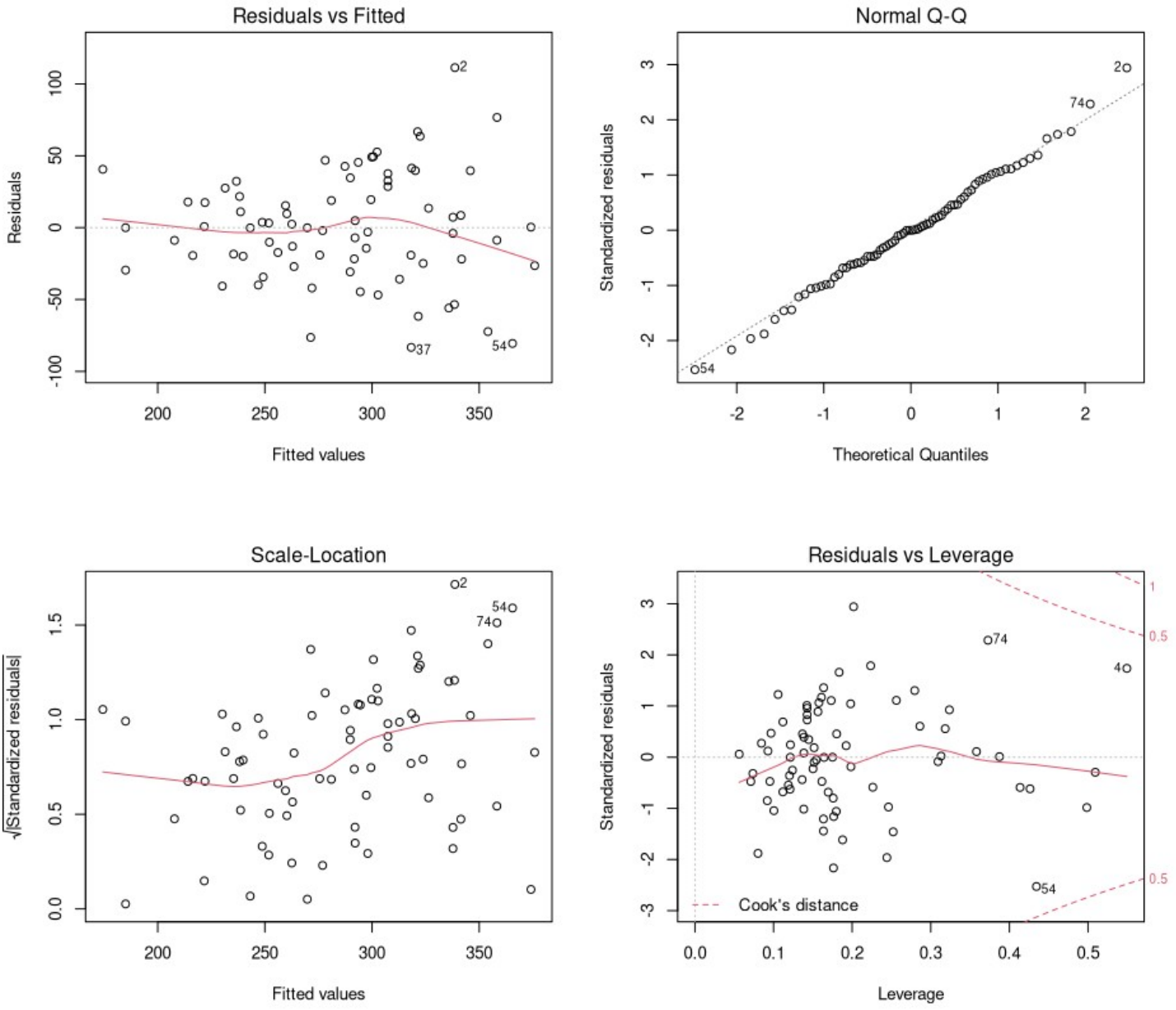
ค่าสัมประสิทธิ์การถดถอยของ Active มีความหมายว่า บ้านที่อยู่ในรายการขายจะมีราคามากกว่าบ้านที่ขายแล้วอยู่ 30.3595 หน่วย

ค่าสัมประสิทธิ์การถดถอยของ DHa มีความหมายว่า บ้านที่อยู่ใกล้โรงเรียน Harris จะมีราคามากกว่า บ้านที่อยู่ใกล้โรงเรียน Redwood อยู่ 40.2357 หน่วย

3.3 จงตรวจสอบข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยด้วยแผนภาพของส่วนเหลือ (residual plot) และอธิบายลักษณะของแผนภาพว่า มีความสอดคล้องกับข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอย หรือไม่อย่างไร

คำสั่ง R

```
par(mfrow=c(2,2))
plot(model)
```



จากแผนภาพ

Residual vs fitted plot : ไม่สอดคล้อง เพราะ ค่าเฉลี่ยความคลาดเคลื่อนมีค่าไม่เท่ากับ 0

Normal Q-Q plot : สอดคล้อง เพราะ การกระจายของความคลาดเคลื่อนเป็นการแจกแจงปกติ

Scale-location plot : ไม่สอดคล้อง เพราะ ค่าคลาดเคลื่อนไม่คงที่

Residual vs leverage plot : สอดคล้อง เพราะ ไม่มีค่านอกกลุ่ม

ข้อ 4. จงหาสมการถดถอยที่เหมาะสมสำหรับการทำนายราคาขายบ้าน โดยที่ตัวแปรทำนายในสมการถดถอยดังกล่าวมีความสัมพันธ์กับราคาขายบ้านอย่างมีนัยสำคัญทางสถิติทุกตัวแปร (พิจารณา p-value ที่ระดับนัยสำคัญ 0.1, 0.05, 0.01 หรือ 0.001)

ตัวแบบการถดถอย คือ

$$Price = \beta_0 + \beta_1 Floor + \beta_2 Lot + \beta_3 Bath + \beta_4 Bed + \beta_5 Bath * Bed + \beta_6 Age^2 + \beta_7 Active + \beta_8 D_{Ed} + \beta_9 D_{Ha} + \varepsilon_i$$

คำสั่ง R

```
model <- lm(Price ~ Floor + Lot + Bath + Bed + I(Bath*Bed) + I(Age^2) + Active +
    DEd + DHa, data = Homes)
```

จากผลลัพธ์ของคำสั่ง R ข้างต้น

สมการถดถอยพหุคูณของตัวอย่าง คือ

$$Price = 319.085 + 71.360\, Floor + 10.616\, Lot - 82.133\, Bath - 81.999\, Bed + 27.522\, Bath * Bed + 1.237\, Age^2 + 31.855\, Active + 62.787\, D_{Ed} + 51.585\, D_{Ha} + \varepsilon_i$$

R^2 มีค่าเท่ากับ 0.553 หมายถึง สมการเชิงเส้นนี้สามารถอธิบายความแปรปรวนของ Price ได้ 0.553

ความคลาดเคลื่อนมาตรฐานของส่วนเหลือ (Residual standard error) มีค่าเท่ากับ 43

ค่าสถิติทดสอบ F มีค่าเท่ากับ 9.072 ค่า p-value เท่ากับ 8.816e-09 ซึ่งสรุปได้ว่า มีค่า β ตัวใดตัวหนึ่ง ไม่เท่ากับ 0

การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอยของสมการถดถอย

สมมติฐานการทดสอบ คือ $H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$

ค่าสถิติทดสอบ

ตัวแปร	ค่าประมาณสัมประสิทธิ์การถดถอย $\hat{\beta}_j$	ความคลาดเคลื่อนมาตรฐานของ $\hat{\beta}_j$ $SE(\hat{\beta}_j)$	สถิติทดสอบที t-statistic	p-value
Intercept	319.085	108.943	2.929	0.004665
Floor	71.360	27.978	2.551	0.013083
Lot	10.616	3.416	3.108	0.002777
Bath	-82.133	43.148	-1.903	0.061338
Bed	-81.999	28.443	-2.883	0.005314
Bath*Bed	27.522	12.119	2.271	0.026415
Age^2	1.237	0.722	1.714	0.091268
Active	31.855	11.090	2.872	0.005473
DEd	62.787	16.509	3.803	0.000315
DHa	51.585	14.978	3.444	0.001000

จากผลการทดสอบสมมติฐานข้างต้น จงสรุปว่า ตัวแปรทำนายตัวใดบ้างที่มีผลต่อราคาขายของบ้าน อย่างมีนัยสำคัญทางสถิติ (พิจารณา p-value ที่ระดับนัยสำคัญ 0.1, 0.05, 0.01 หรือ 0.001)

ตัวแปร DEd มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.001

ตัวแปร Intercept, Lot, Bed, Active, DHa มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.01

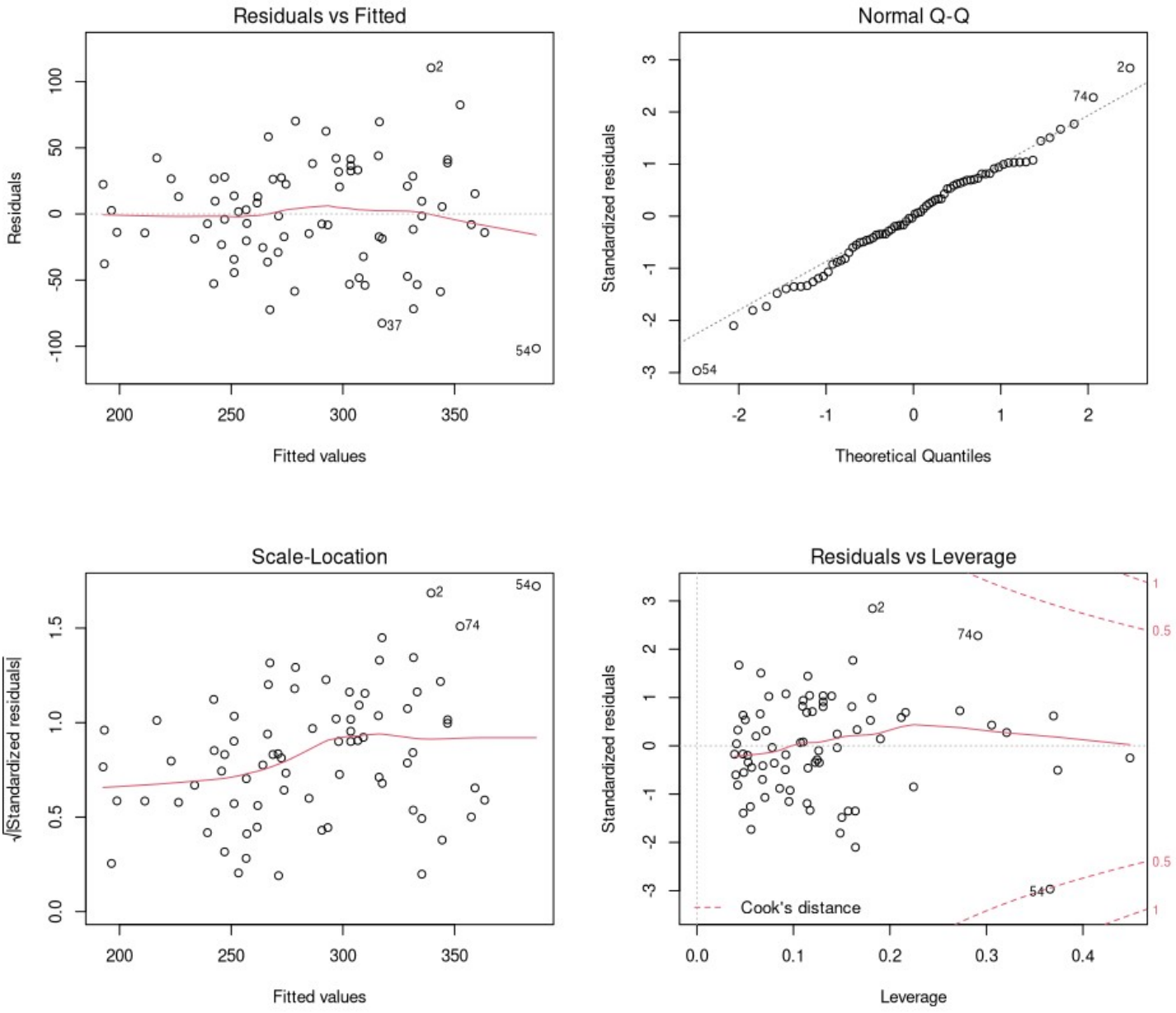
ตัวแปร Floor, Bath * Bed มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05

ตัวแปร Bath, Age^2 มีผลต่อราคาบ้าน อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.1

การตรวจสอบข้อสมมติเบื้องต้นของการวิเคราะห์การถดถอยด้วยแผนภาพของส่วนเหลือ (residual plots)

คำสั่ง R

```
par(mfrow=c(2,2))
plot(model)
```



จงอธิบายลักษณะของแผนภาพว่า มีความสอดคล้องกับข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยหรือไม่อย่างไร

Residual vs fitted plot สอดคล้อง เพราะ ค่าเฉลี่ยความคลาดเคลื่อนมีค่าคงที่ และมีการกระจายคงที่

Normal Q-Q plot ไม่สอดคล้อง เพราะ การกระจายของความคลาดเคลื่อนไม่เป็นการแจกแจงปกติ

Scale-location plot ไม่สอดคล้อง เพราะ ค่าความคลาดเคลื่อนไม่คงที่

Residual vs leverage plot ไม่สอดคล้อง เพราะ มีค่านอกกลุ่ม