Rで学ぶ
単回帰分析と重回帰分析

M2 新屋裕太

2013/05/29

発表の流れ

1.回帰分析とは？

2.単回帰分析

単回帰分析とは？　/　単回帰式の算出　/　単回帰式の予測精度

＜Rによる演習①＞

3.重回帰分析

重回帰分析とは？　/　重回帰式の算出　/　重回帰式の予測精度

質的変数を含む場合の回帰分析　/　多重共線性の問題

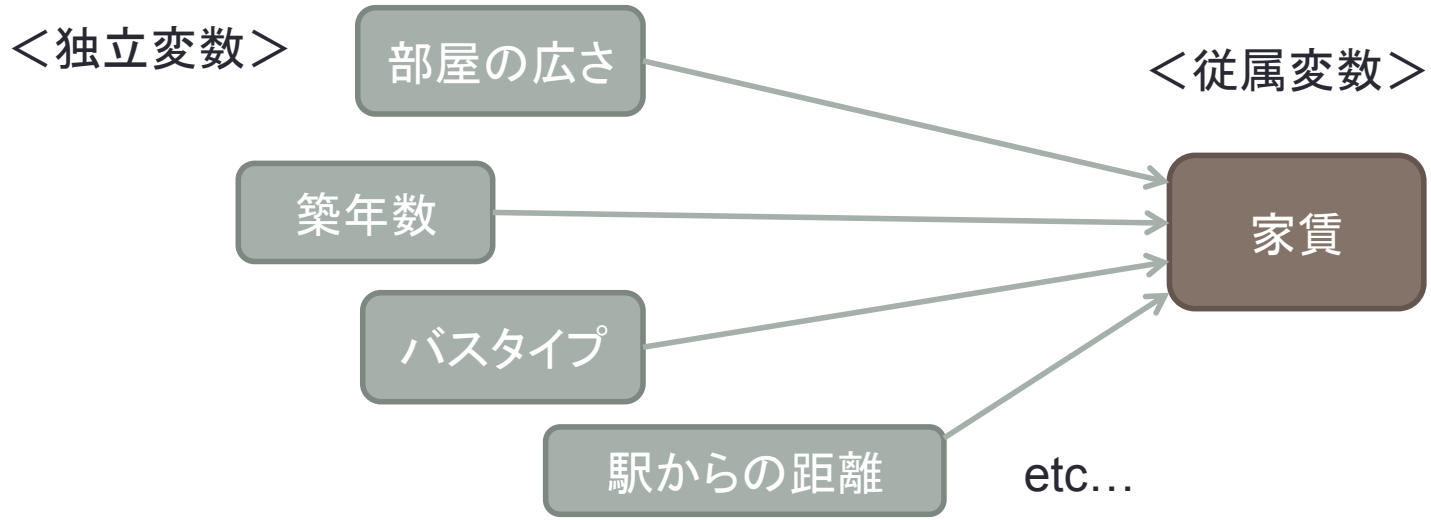
変数選択の基準と方法

＜Rによる演習②＞

回帰分析とは？

- 変数間の因果関係の方向性を仮定し、1つまたは複数の独立変数による従属変数の予測の大きさ（説明率）を検討する分析
  - 単回帰分析：予測変数が1つの場合
  - **重回帰分析**：予測変数が2つ以上の場合

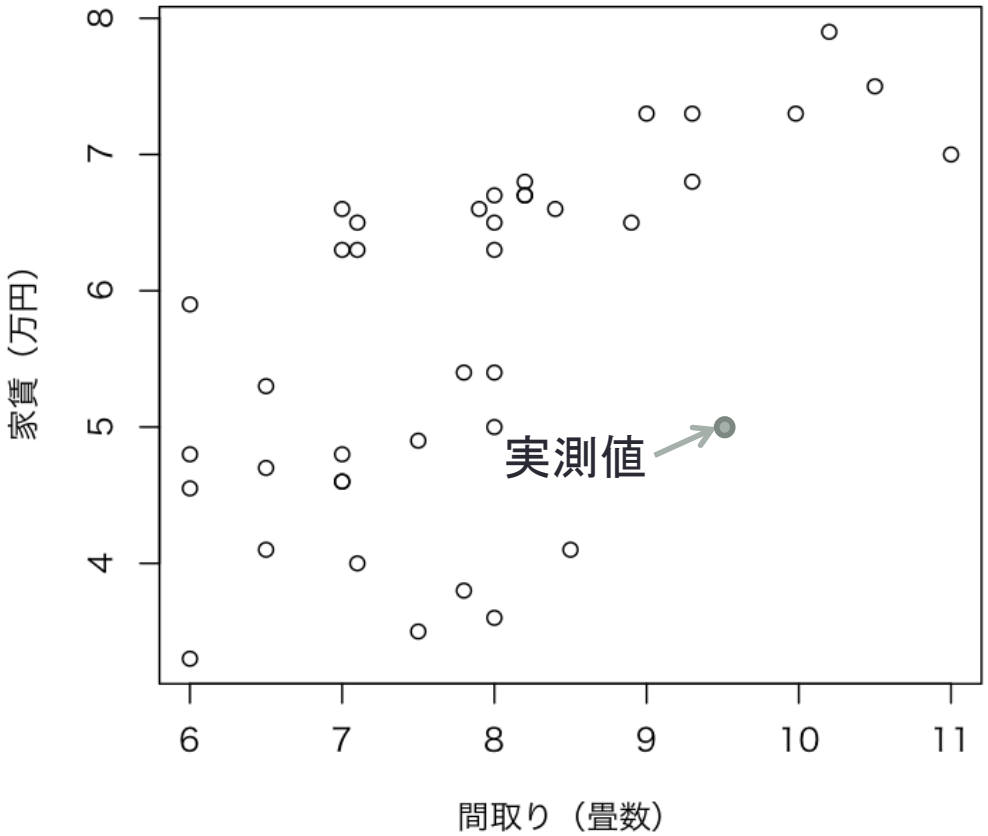
（例）ワンルームマンションの家賃を、ワンルームマンションの条件から、予測する場合



単回帰分析とは？

- 単回帰分析では、独立変数xと従属変数yの間に、以下のような線的の関係があることを仮定する
- y = a + bx + e（単回帰モデル）
- y^ = a + bx（単回帰式）
  - y : 実測値
  - y^: 予測値
  - a : 切片
  - b : 傾き（回帰係数）
  - e : 誤差（残差）

（例）吉田キャンパス周辺のワンルームマンションの家賃を予測する場合



単回帰式の算出

- 実際のデータ、実測値yは、あるxに対してさまざまな値をとりうる

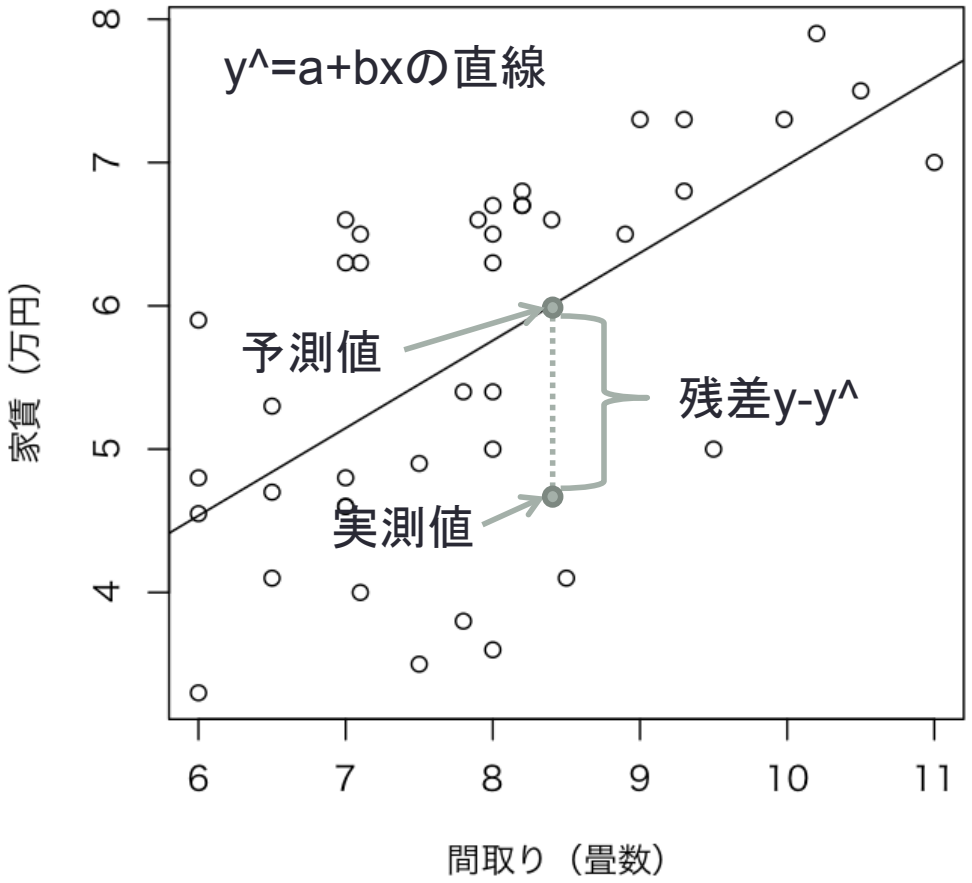
→残差（実測値−予測値）の最も少ない回帰式を求めたい

最小2乗法によって、誤差（残差）の平方和が最小になるような定数項a,bを求める

誤差平方和:

$$Q = \sum[y_i - (a+bx_i)]^2$$

・・・aとbを偏微分し、結果を0とした連立方程式の解によって求められる



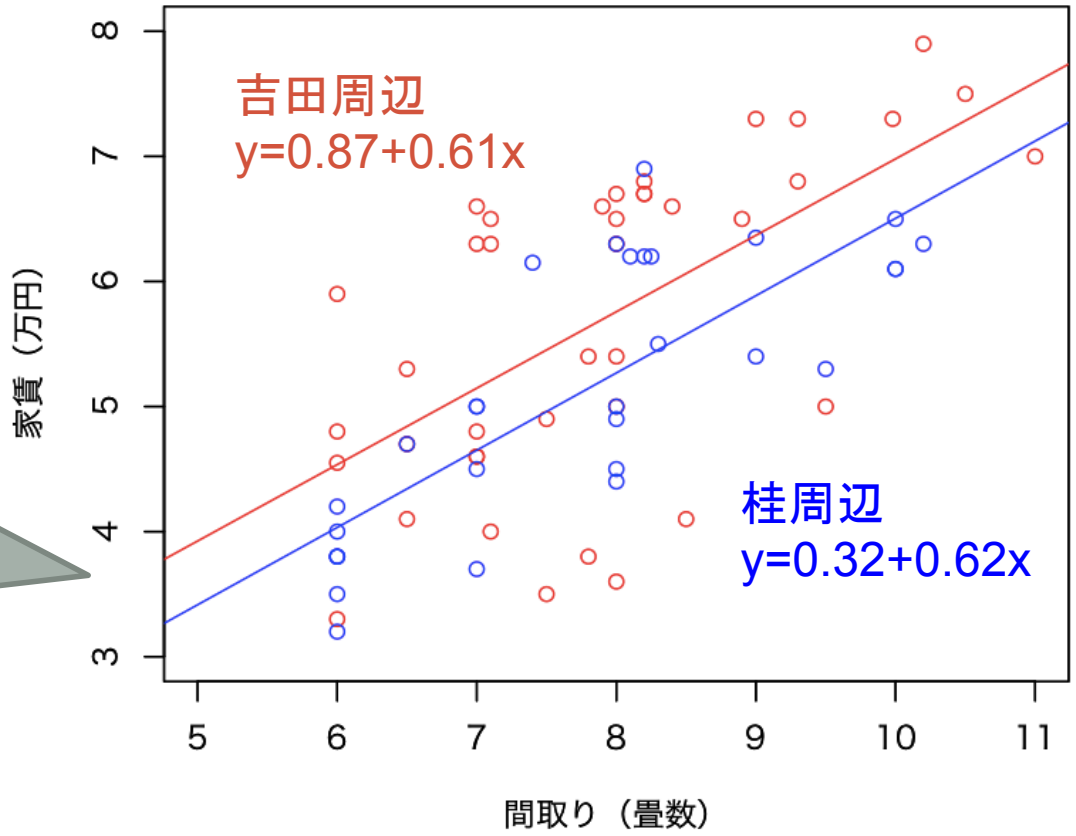
単回帰式の算出

- 得られた単回帰式：y^=0.87+0.61x
  （例）6.5帖の場合　y^=0.87+0.61×6.5=4.835（万円）

→みなさんの下宿はどうでしょうか？

- ちなみに、桂周辺だと、、、
- 単回帰式：
- y^ = 0.32+0.62x

傾きはほどんど同じだが、切片が5000円以上異なる

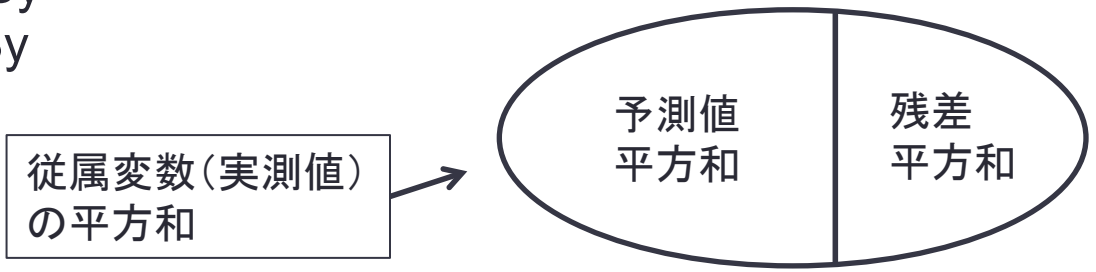


単回帰式の予測精度

- 回帰式によって得られた予測値は、どれくらい実測値を予測しているのか？
  - 残差の平方和（分散）を残差の大きさとして予測の精度を測る

- 回帰式の精度を表す指標
  - SSy（実測値の平方和）=SSy^（予測値の平方和）+SSe（残差の平方和）
  - 1 = SSy^/SSy + SSe/SSy
  - SSy^/SSy = 1 - SSe/SSy

決定係数(R^2)

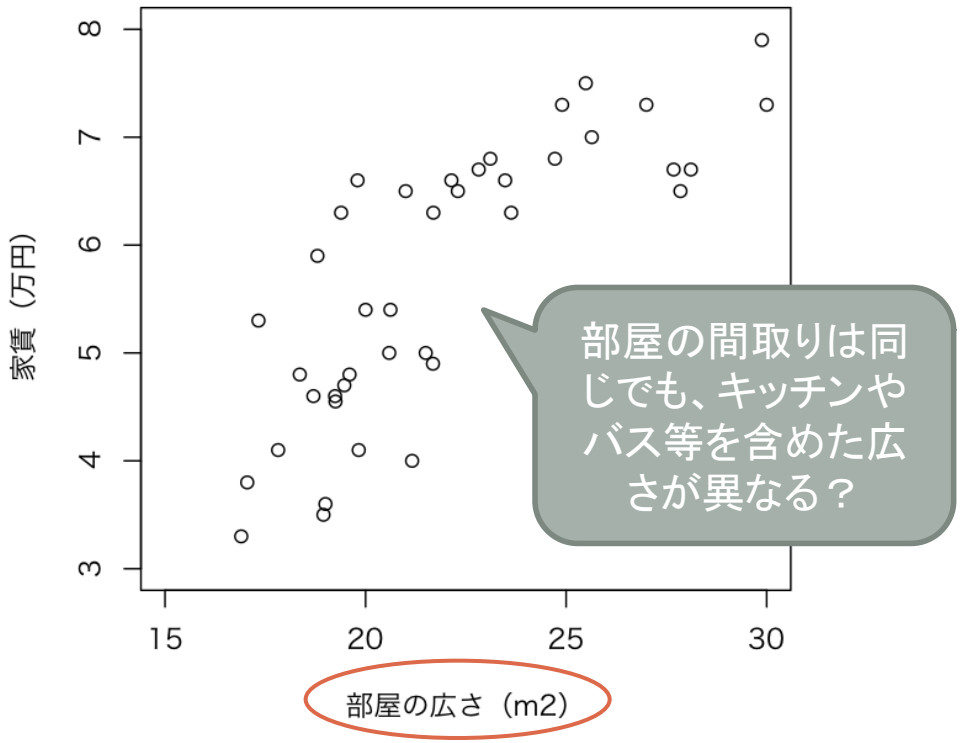


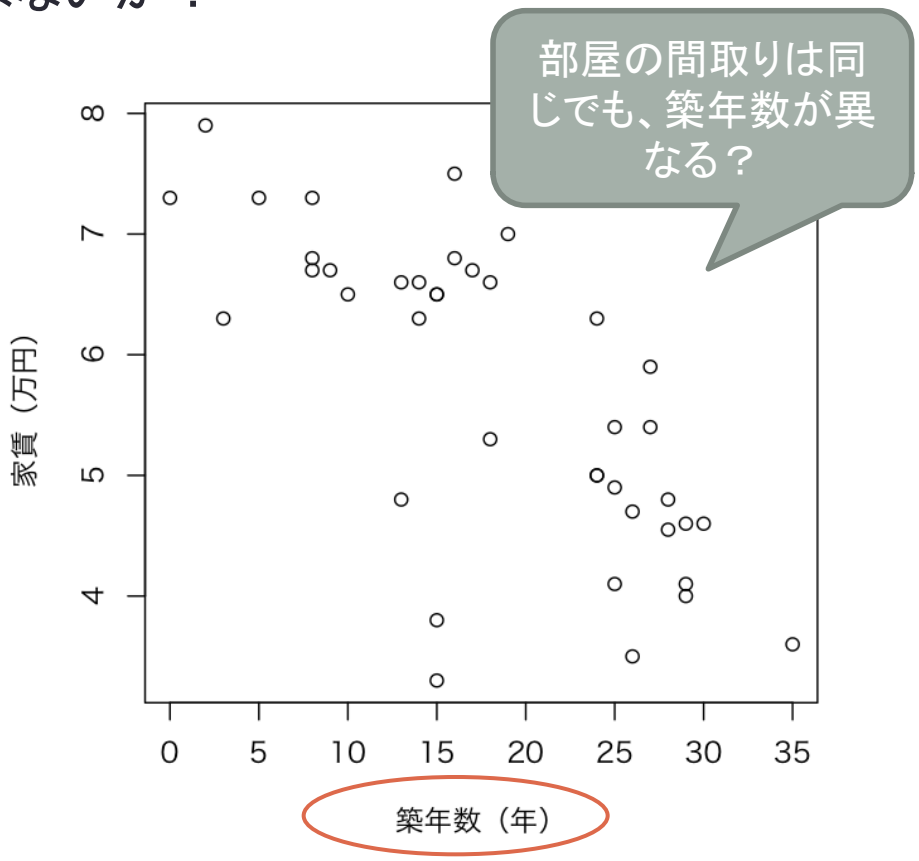
- 決定係数（R^2）は説明変数によって説明される分散の割合を示す
  1に近いほど予測の精度が高い
  - 決定係数が0に近いほど円状の分布、1に近いほど回帰直線に近似する分布を取る

単回帰式の予測精度

- ワンルームマンションの例（部屋の間取り→家賃）だと、
  - R^2=0.3673(分散の約36.7%を説明)

  ・・・もう少し予測の精度が高い変数はないか？





Rによる演習①

＝＞他の変数と従属変数の単回帰式・予測の精度を求めてみよう

- **部屋の広さ(m^2)**→家賃
  - 単回帰式：
  - R^2=
- **築年数**→家賃
  - 単回帰式：
  - R^2=

Rによる演習①

- 分析の下準備
  - R Consoleを起動
  - 「ファイル」→「ディレクトリの変更」で、data.csvが保存されているフォルダを選ぶ
  - 「ファイル」→「新しいスクリプト」を選び、スクリプトエディタを開く
  - 実行したい作業・分析を書き込む→その部分を選択し、Ctrl+R（Macの場合はcommand+enter）で実行する
  - 結果はR Consoleに表示される

  - データの読み込み
    - dat<-read.csv(“data0.csv”)
  - データ範囲の絞り込み（zone 0が吉田、1が中京、2が桂です）
    - dat0<-subset(dat,zone=="0")

Rによる演習①

- データの確認
  - dat0

zone:地域、rent:家賃
area1：間取り（帖数）、area2:広さ（m^2）、age:築年、bath:バスタイプ

	zone	rent	area1	area2	age	bath
1	0	7.30	9.00	27.00	0	0
2	0	6.50	8.00	27.85	15	0
3	0	7.90	10.20	29.88	2	0
4	0	5.90	6.00	18.80	27	0
5	0	6.80	8.20	24.72	8	0
6	0	5.40	8.00	20.62	25	1
7	0	5.30	6.50	17.33	18	1
8	0	7.00	11.00	25.64	19	0
9	0	7.50	10.50	25.49	16	0
10	0	4.80	6.00	18.36	13	1
11	0	6.30	7.10	23.63	3	0
12	0	6.60	7.00	19.80	14	0
13	0	7.30	9.30	24.90	8	0
14	0	6.70	8.20	22.82	17	0
15	0	6.50	8.90	22.30	15	0
16	0	4.60	7.00	18.70	30	1
17	0	7.30	9.98	30.00	5	0
18	0	3.30	6.00	16.90	15	1
19	0	6.70	8.00	27.68	8	0
20	0	6.80	9.30	23.11	16	0

21	0	3.80	7.80	17.05	15	1
22	0	5.00	9.50	20.59	24	1
23	0	6.30	8.00	21.69	24	1
24	0	6.60	8.40	23.48	18	0
25	0	6.70	8.20	28.11	9	0
26	0	6.60	7.90	22.14	13	0
27	0	5.00	8.00	21.50	24	1
28	0	6.30	7.00	19.39	14	0
29	0	6.50	7.10	21.00	10	0
30	0	4.90	7.50	21.68	25	1
31	0	4.80	7.00	19.60	28	1
32	0	5.40	7.80	20.00	27	1
33	0	4.60	7.00	19.24	29	1
34	0	3.50	7.50	18.95	26	1
35	0	4.10	8.50	17.82	29	1
36	0	4.10	6.50	19.83	25	1
37	0	3.60	8.00	19.00	35	1
38	0	4.70	6.50	19.47	26	1
39	0	4.55	6.00	19.25	28	1
40	0	4.00	7.10	21.16	29	1

Rによる演習①

- 回帰分析（lm関数を使用）
  - lm(rent~age,data=dat0)

従属変数　説明変数　　参照データ、ここではdat0を指定

→切片（intercept）、回帰係数が算出される

- 決定係数を含む詳細な結果
  - reg1<-lm(rent~age,data=dat0)
  - summary(reg1)

```
> summary(lm(rent~age,data=dat0))

Call:
lm(formula = rent ~ age, data = dat0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.72841 -0.29669  0.05258  0.48708  1.57452

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.57230    0.32803  23.084  < 2e-16 ***
age         -0.10293    0.01616  -6.369 1.78e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8975 on 38 degrees of freedom
Multiple R-squared:  0.5163,    Adjusted R-squared:  0.5036
F-statistic: 40.56 on 1 and 38 DF,  p-value: 1.778e-07
```

残差分布の四分位数

切片・傾きの推定値と検定結果

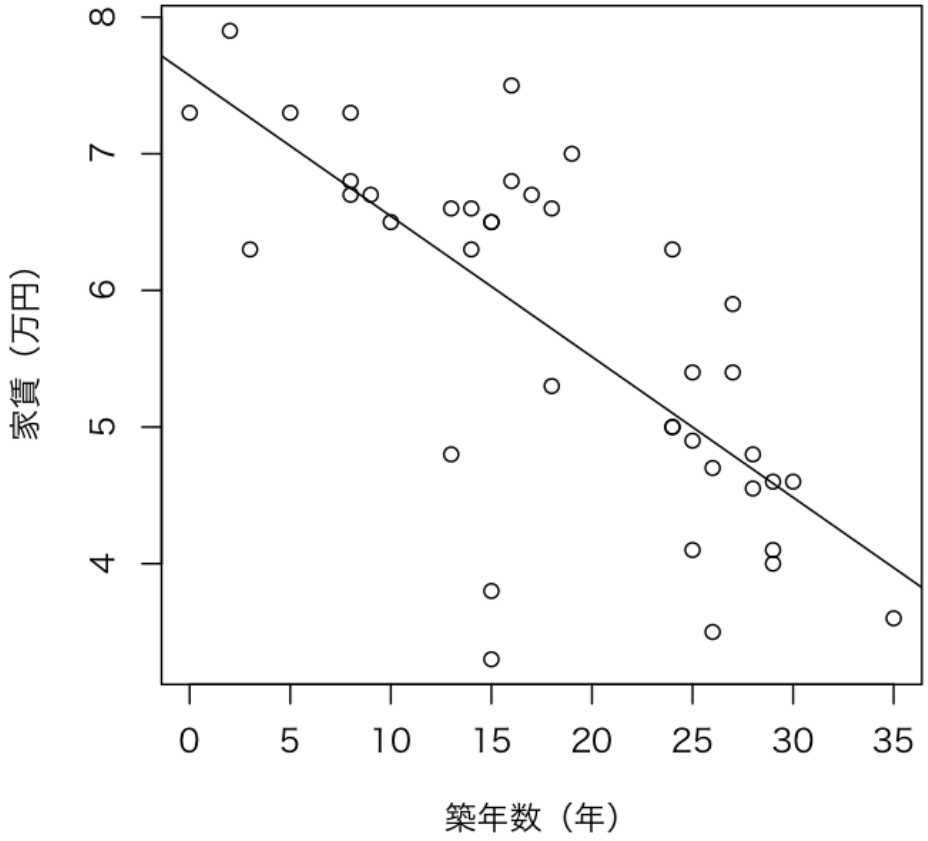
決定係数

Rによる演習①

- 散布図を描く
  - plot(dat0$age,dat0$rent,xlab="築年数（年）",ylab="家賃（万円）")

X軸（説明変数）　Y軸（従属変数）　X軸・Y軸のラベル

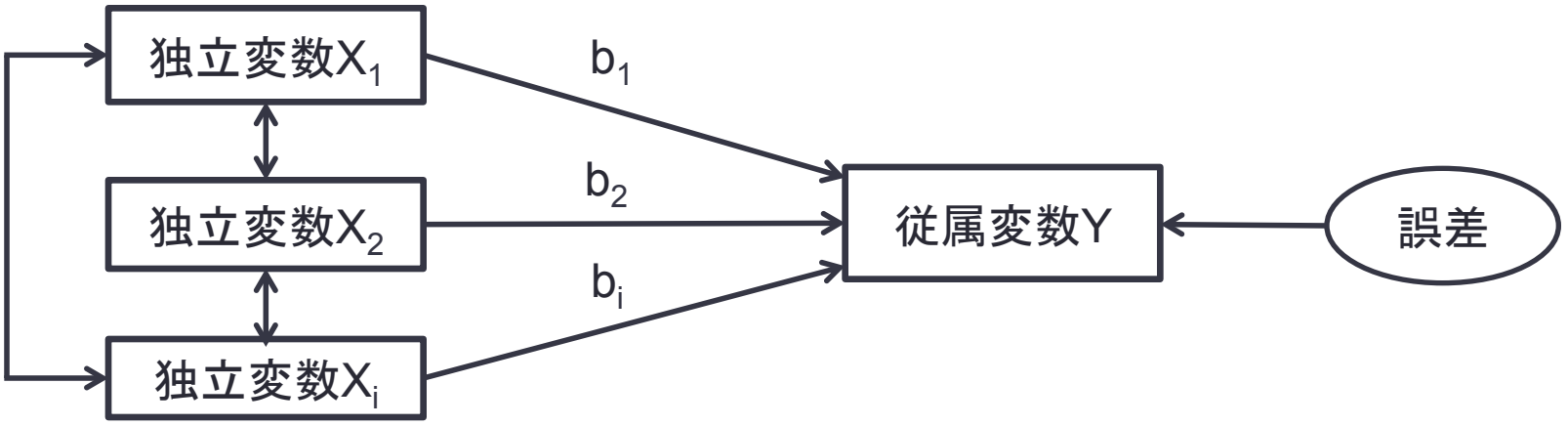
- 単回帰直線を描く
  - abline(reg1)



重回帰分析とは？

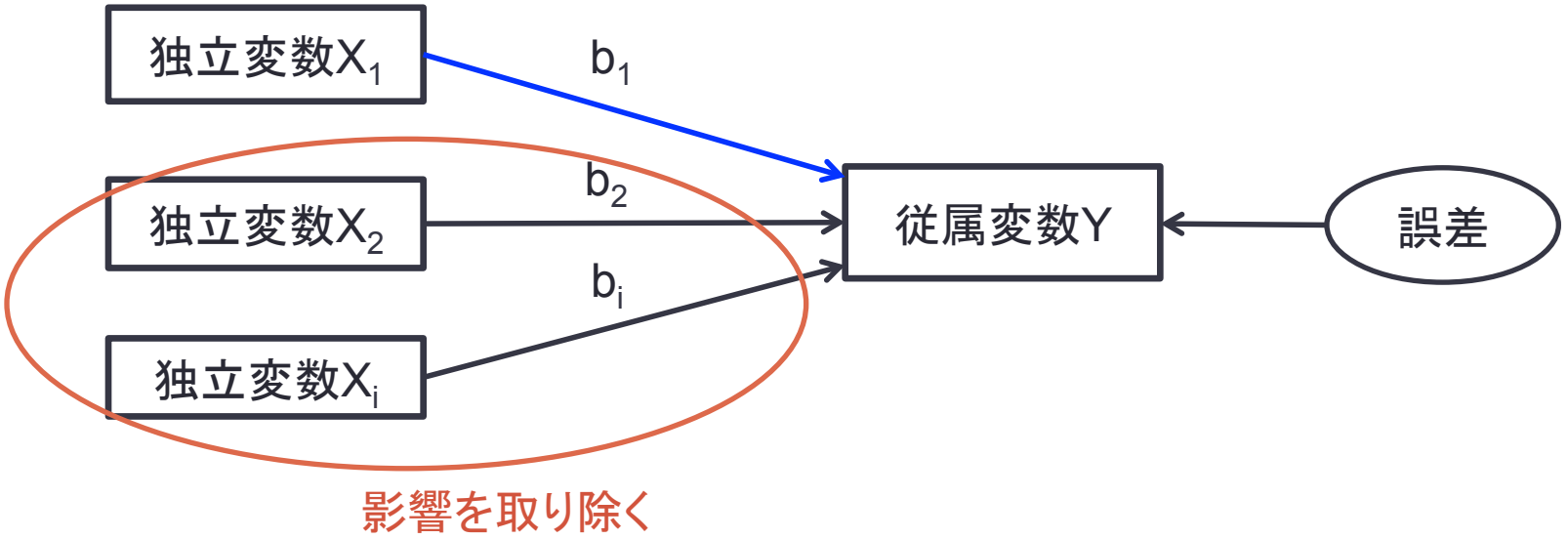
- 重回帰分析では、複数個の独立変数$x_1, x_2, \cdots, x_i$と従属変数yの間に、以下のような線形の関係があることを仮定する

- $y = a + b_1x_1 + b_2x_2 + \cdots + b_ix_i + e$（重回帰モデル）
- $y\hat{} = a + b_1x_1 + b_2x_2 + \cdots + b_ix_i$（重回帰式）
  - y^:予測値　a:切片　b:偏回帰係数　e: 誤差（残差）



偏回帰係数

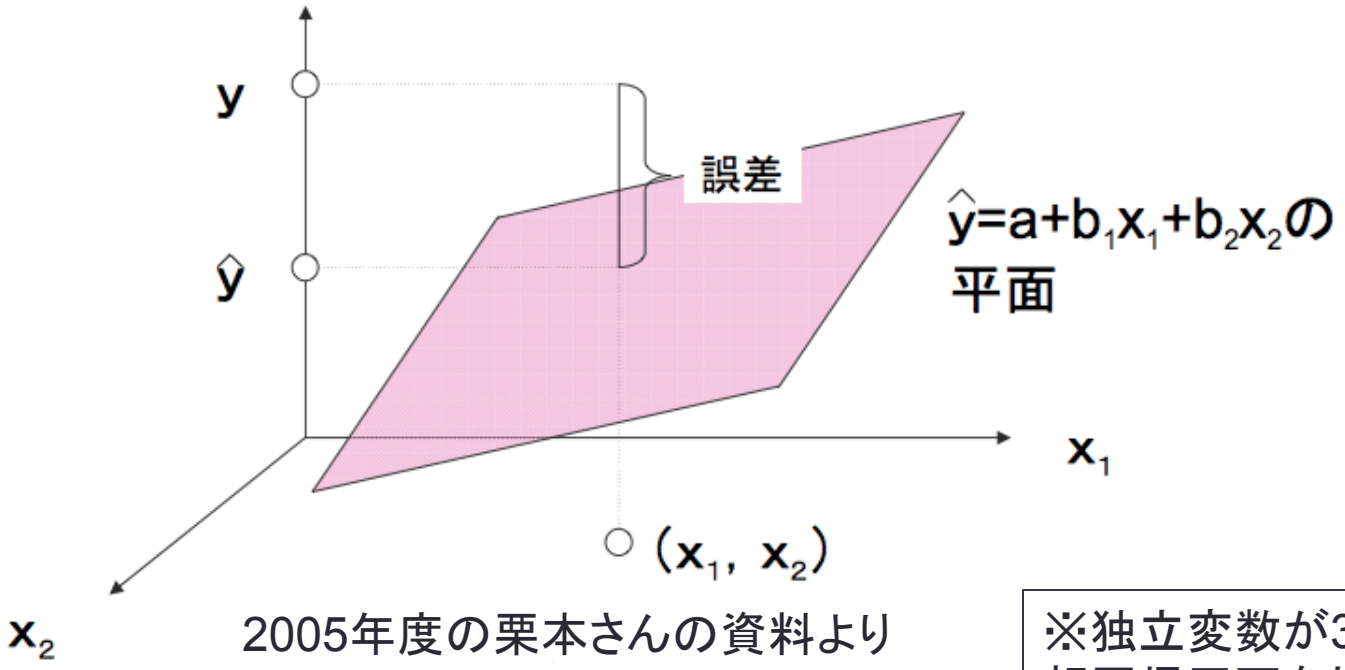
- 偏回帰係数は他の独立変数の影響を除いた上で、ある独立変数の値が1変わった時に従属変数の値が平均的にどれだけ変化するかを示す



- 偏回帰係数は、独立変数・従属変数の単位に依存するため、単位やスケールが異なる場合は標準化する
  - 標準偏回帰係数=偏回帰係数×（独立変数のSD/従属変数のSD）

重回帰式の算出

- 単回帰分析の場合と同じく、最小2乗法によって、残差の2乗和が最も少なくなるような切片（a）と偏回帰係数（b）を求める
  - 3変数の回帰式 $y\hat{}=a+b_1x_1+b_2x_2$は平面を表す



2005年度の栗本さんの資料より

※独立変数が3つ以上の場合は、超回帰平面をとる

重回帰式の予測精度

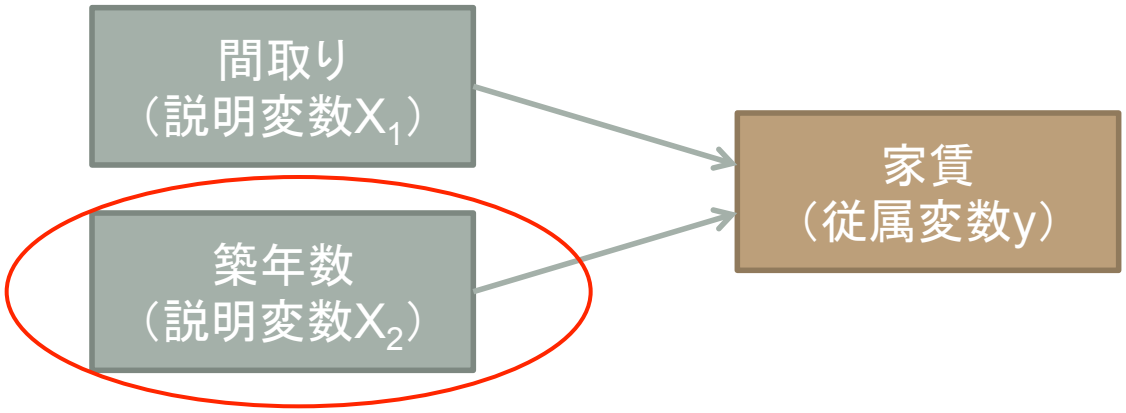
- 単回帰の場合と同じく、残差の分散を残差の大きさとして予測の精度を測る

- 回帰式の精度を表す指標
  - 重相関係数（R）
    - 予測変数y^と従属変数yの相関係数
  - 決定係数(R^2)
    - SSy（従属変数の分散）=SSy^（予測値の分散）+SSe（誤差の分散）
    - 両辺をSSyで割ると、1 = SSy^/SSy + SSe/SSy
    - 決定係数（もしくは分散説明率）: SSy^/SSy=1-SSe/SSy
  - 自由度調整済み決定係数（R^{*2}）
    - 独立変数の数を考慮したモデル
    - R^{*2}=1- SSe/n-k-1 / SSy/n-1
    - n:サンプル数　k:独立変数の数

重回帰式の予測精度

- 決定係数（R^2）は説明変数によって説明される分散の割合を示す、1に近いほど予測の精度が高い
  - 決定係数が0に近いほど、球状の分布を取る
  - 決定係数が1に近いほど、回帰平面に近似する分布を取る

- ワンルームマンションの家賃の例：
  →間取り＋築年数から家賃を予測する



重回帰式の予測精度

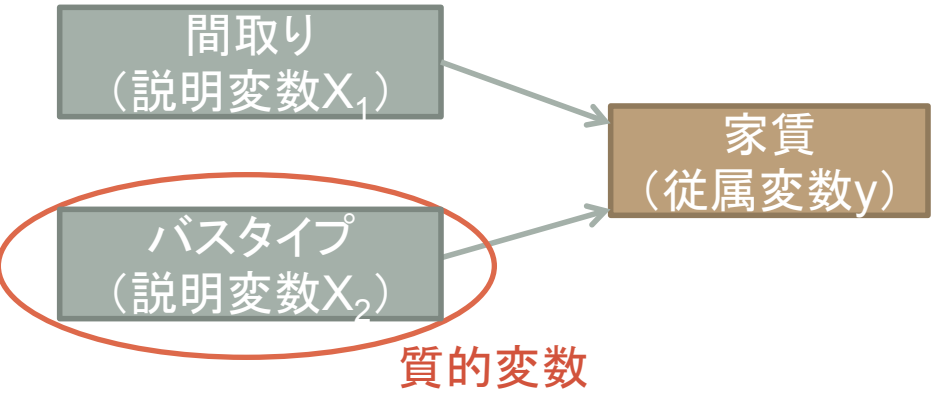
- 重回帰式：$y = 3.99 + 0.41x_1 - 0.08x_2$
- （標準化した場合：$y=0.40x_1 - 0.58x_2$）

→決定係数：$R^2 = 0.66$, $R^{*2} = 0.64$まで上昇！

```
> summary(lm(rent~area1+age,dat0))

Call:
lm(formula = rent ~ area1 + age, data = dat0)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1260 -0.3416  0.1417  0.4504  1.6967

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9951     0.9528   4.193 0.000165 ***
area1         0.4064     0.1035   3.927 0.000362 ***
age          -0.0826     0.0147  -5.618 2.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7642 on 37 degrees of freedom
Multiple R-squared:  0.6586,    Adjusted R-squared:  0.6401
F-statistic: 35.69 on 2 and 37 DF,  p-value: 2.322e-09
```

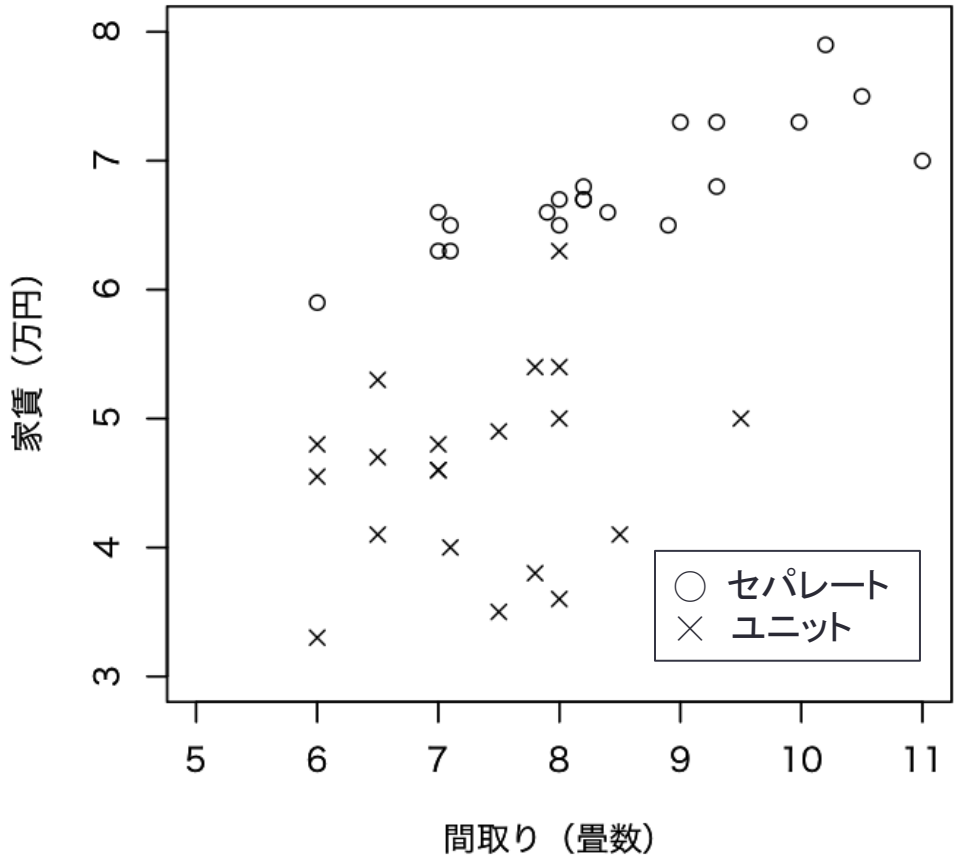
質的変数を含む場合の回帰分析

- 説明変数に質的変数が含まれる回帰分析



→ダミー変数dを利用して、
変数の効果を検討する

- $d = \begin{cases} 0 & \text{セパレートバス} \\ 1 & \text{ユニットバス} \end{cases}$



質的変数を含む場合の回帰分析

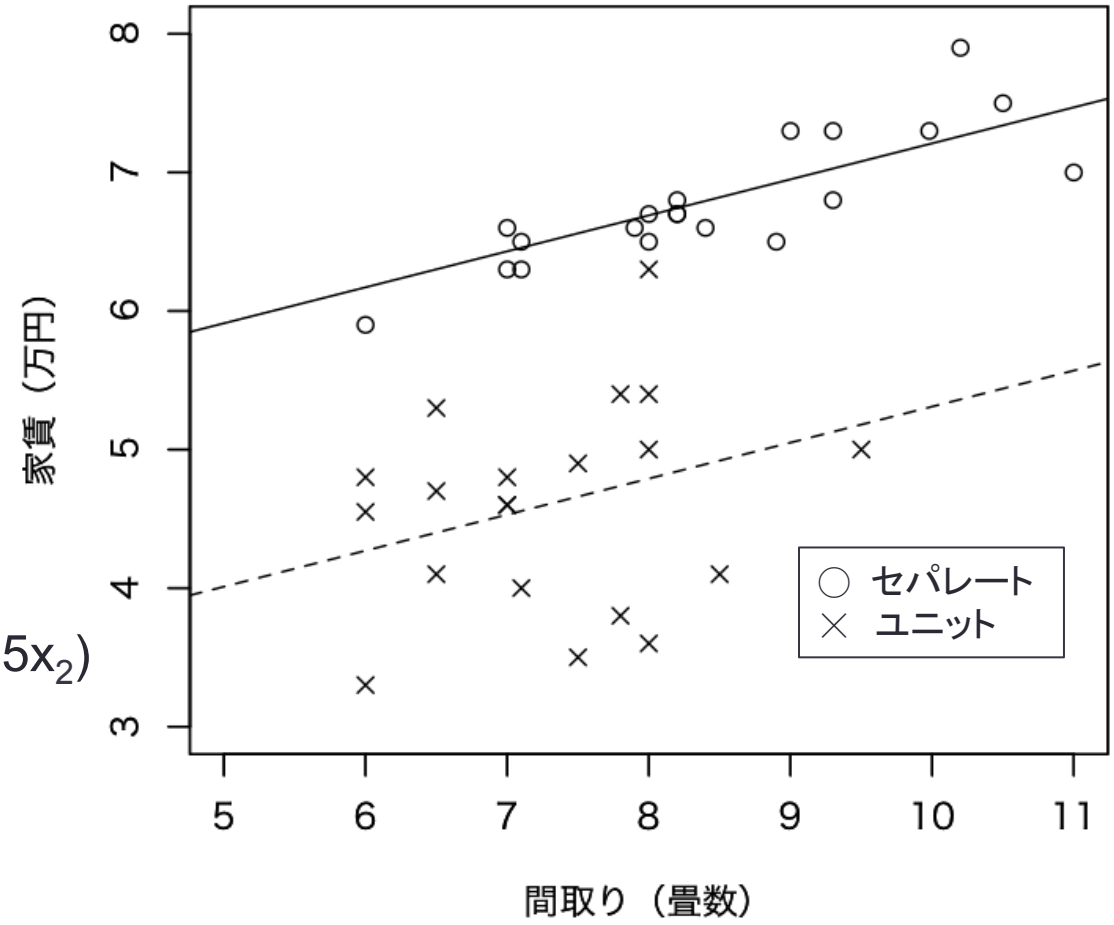
- カテゴリー間で切片が異なる重回帰モデルを以下の式で表現する
  - $Y = a+b_1x_i+\mathbf{b_2d}+e$
- d=0の場合、
  - $Y = a+b_1x_i+e$
- d=1の場合
  - $Y = (a+b_2)+b_1x_i+e$
  - と表される

- 重回帰式:
  - $y = 4.61 + 0.26x_1 - 1.90x_2$

(標準化した場合:$y=0.26x_1-0.75x_2$)

→決定係数:$R^2 = 0.82$

$R^{*2} = 0.81$



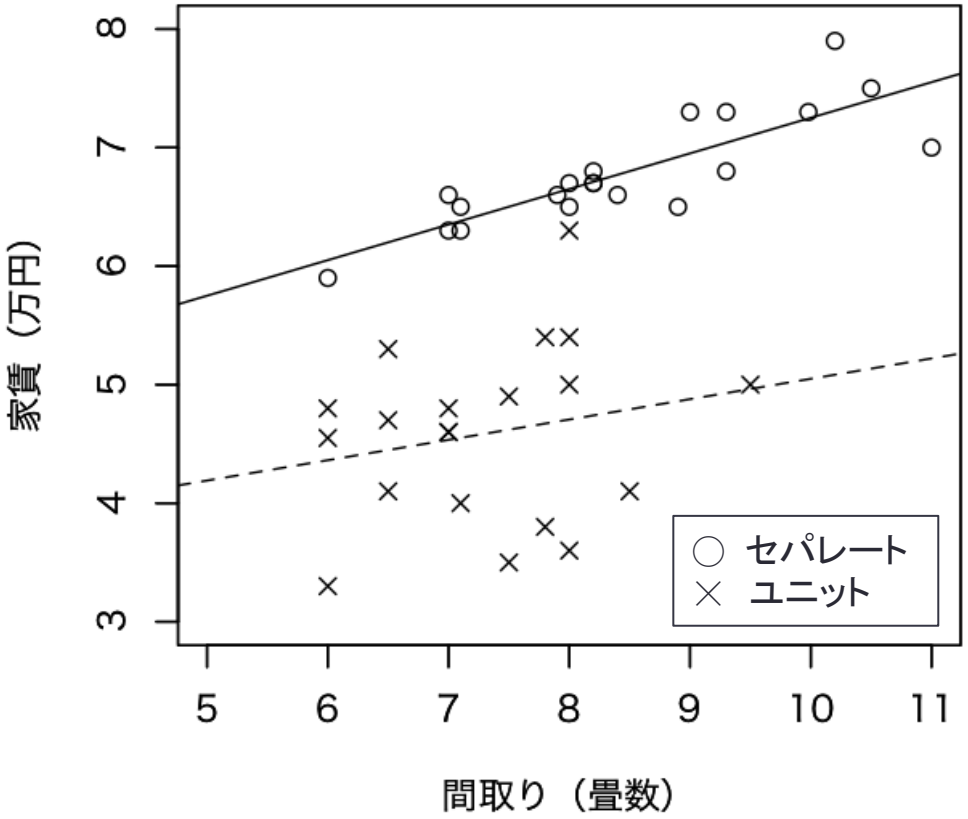
質的変数を含む場合の回帰分析

- ただし、実際にはカテゴリー間で切片だけでなく傾きも異なる可能性があるのでは？
- ある独立変数の効果が他の独立変数によって異なる

→交互作用の検討

- 重回帰分析においても交互作用の検討が可能

→次回の発表で取り扱います！



多重共線性の問題

- 独立変数間の相関が高すぎる場合には偏回帰係数の推定量が不安定になる。（e.g. 係数の絶対値や標準誤差が非常に大きい、係数の符号が実態に則さないなど）

- 相関の強い独立変数を取り除くか、新しい変数を加えるか、相関する複数の変数を一つの変数に合成するなどの方法をとる必要。

- VIF（Variance Inflation Factor，分散拡大要因）
  - VIF=1/(1-Rj)
    - Rj:変数xjを従属変数、他の変数を独立変数にしたときの決定係数
  - 多重共線性が生じているかどうかを判断する指標
    - VIF>10であれば、可能性を疑うべき

変数選択の基準と方法

- 一度に多くの予測変数を利用すると、多重共線性などの問題が生じる可能性も高くなる
  - 有効な予測変数のみを選択して、精度の高い重回帰モデルを構築する必要

- 変数選択の基準
  - 自由度調整済決定係数（R^{*2}）
    - 誤差分散を誤差の自由度で、分散全体を全体の自由度で割る

    →値が高いほどよいモデルとみなす
  - AIC（Akaike's Informataion Criteriaon, 赤池情報量基準）
    - データとモデルの当てはまりの良さを測る指標

    →値が小さいほどよいモデルとみなす

- 変数選択の方法
  - 総当り法：予測変数の候補がp個の場合、2^p-1個の回帰式を推定し比較
  - 逐次選択法：特定の基準を元に変数を逐次的に追加・削除する方法
    - 変数増加法、変数減少法、ステップワイズ法

Rによる実習②

=>実際に重回帰分析（説明変数は4つ）を行い、従属変数をよりよく説明できる重回帰式を求めてみよう

①逐次選択法（ステップワイズ法）による変数の選択

- reg0<-lm(rent~1,dat0)　切片のみのモデル
- step(reg0,direction="both",　変数増加法の場合は"forward"

scope=list(upper=~area1+area2+age+bath))　今回は4つの説明変数から選ぶ

Rによる実習②

- 出力結果

```
Start:  AIC=20.35                    切片のみのAIC
rent ~ 1                             （初期値）

        Df Sum of Sq    RSS     AIC
+ bath   1    48.510 14.777 -35.831
+ area2  1    40.000 23.287 -17.639
+ age    1    32.675 30.612  -6.699
+ area1  1    23.246 40.041   4.041
<none>                63.287  20.352
```

各変数を足した場合のAIC

最もAICが低下するbathを選択

```
Step:  AIC=-35.83
rent ~ bath

        Df Sum of Sq    RSS     AIC
+ area2  1     4.095 10.682 -46.813
+ area1  1     3.256 11.522 -43.786
+ age    1     0.783 13.995 -36.008
<none>                14.777 -35.831
- bath   1    48.510 63.287  20.352
```

引いた場合のAIC

area2を選択

```
Step:  AIC=-46.81
rent ~ bath + area2

        Df Sum of Sq    RSS     AIC
+ area1  1    0.6140 10.068 -47.181
<none>                10.682 -46.813
+ age    1    0.0804 10.602 -45.115
- area2  1    4.0954 14.777 -35.831
- bath   1   12.6051 23.287 -17.639
```

area1を選択

```
Step:  AIC=-47.18
rent ~ bath + area2 + area1

        Df Sum of Sq     RSS     AIC
<none>                10.0680 -47.181
- area1  1    0.6140 10.6820 -46.813
+ age    1    0.1999  9.8681 -45.983
- area2  1    1.4538 11.5218 -43.786
- bath   1   12.7960 22.8640 -16.373

Call:
lm(formula = rent ~ bath + area2 + area1, data = dat0)

Coefficients:
(Intercept)         bath        area2        area1
    3.37633     -1.58681      0.09247      0.13690
```

最もAICの低い（=当てはまりの良い）モデル

Rによる実習②

②重回帰分析

＜バスタイプ+間取り+広さ→家賃＞

- reg1<-lm(rent~bath+area1+area2, data=dat0)
- summary(reg1)

```
Call:
lm(formula = rent ~ bath + area1 + area2, data = dat0)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06848 -0.27252  0.07612  0.22522  1.40972

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.37633    0.84538   3.994 0.000307 ***
bath        -1.58681    0.23459  -6.764 6.71e-08 ***
area1        0.13690    0.09239   1.482 0.147118
area2        0.09247    0.04056   2.280 0.028637 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5288 on 36 degrees of freedom
Multiple R-squared:  0.8409,    Adjusted R-squared:  0.8277
F-statistic: 63.43 on 3 and 36 DF,  p-value: 1.918e-14
```

area1は帰無仮説（係数=0）を棄却できない
→area1は除く

Rによる実習②

＜広さ+バスタイプ→家賃＞

- reg2<-lm(rent~bath+area2, data=dat0)
- summary(reg2)

```
Call:
lm(formula = rent ~ area2 + bath, data = dat0)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02565 -0.31007  0.07312  0.21037  1.42636

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.69278    0.83107   4.443 7.75e-05 ***
area2        0.12700    0.03372   3.766 0.000576 ***
bath        -1.57384    0.23818  -6.608 9.50e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5373 on 37 degrees of freedom
Multiple R-squared:  0.8312,    Adjusted R-squared:  0.8221
F-statistic: 91.11 on 2 and 37 DF,  p-value: 5.078e-15
```

係数は全て有意
R*2も非常に高い
値を得ることが出
来た

Rによる実習②

どちらの係数もVIF<10であるため、多重共線性は生じていないと判断

③多重共線性の確認

- reg3<-lm(rent~bath+area2,dat0)

```
> #bathのVIFを求める
> rs1<-summary(lm(bath~area2,dat0))$r.squared    #area2→bathの決定係数
> 1/(1-rs1)
[1] 1.965043
> #area2のVIFを求める
> rs2<-summary(lm(area2~bath,dat0))$r.squared  #bath→area2の決定係数
> 1/(1-rs2)
[1] 1.965043
```

- 最終的に得られた重回帰式：$y^{\wedge}=3.69-1.6x_1+0.13x_2$（$R^{*2}=0.82$）

  x1:バスタイプ<0=セパレート 1=ユニット>, x2:部屋の広さ（m^2）

・・・ただ、部屋の広さ（m^2）は把握してない人も多いと思うので、

- 重回帰式：$y = 4.61 + 0.26x_1 – 1.90x_2$（$R^{*2} = 0.81$）

  x1:バスタイプ<0=セパレート 1=ユニット>, x2:間取り（帖数）

  ↑のモデルのほうが使用しやすいかもしれません！

Rによる実習②

もし時間があれば計算してみて下さい！

- 標準回帰係数
  - z <- scale(dat0)　　# 得点を標準化
  - z <- data.frame(z) # データフレーム形式に戻す
  - summary(lm(rent~bath+area2, z))

- 他地域の重回帰式
  - データ範囲の絞り込み
    - 中京：dat1<-subset(dat,zone==“1")
    - 桂：dat2<-subset(dat,zone==“2")
  - 後はdat0→dat1, dat2にして、同様の流れで分析

参考文献

- 南風原朝和（2002）心理学統計の基礎 有斐閣アルマ
- 豊田秀樹（2012）回帰分析入門-Rで学ぶ最新データ- 東京書籍
- 足立浩平（2006）多変量データ解析法 ナカニシヤ出版
- 単回帰分析と重回帰分析（栗本,2005）
- 重回帰分析(魚野;2006) http://kyoumu.educ.kyoto- u.ac.jp/cogpsy/personal/Kusumi/datasem06/uono.pdf
- 重回帰分析(栗田;2008) http://kyoumu.educ.kyoto- u.ac.jp/cogpsy/personal/Kusumi/datasem06/uono.pdf
- 京都ひとり暮らしガイド2013（株）京都住宅センター学生住宅