Rで学ぶ回帰分析

補足:重回帰分析における交互作用の検討

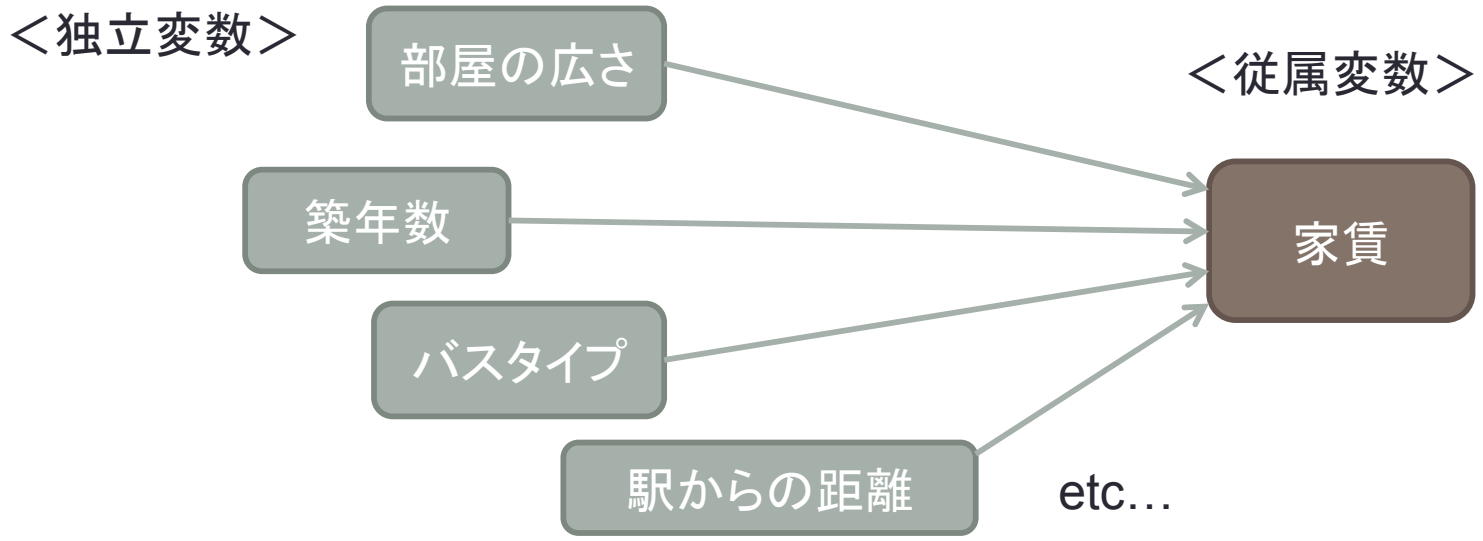
M2 新屋裕太

2013/07/10

（復習）回帰分析について

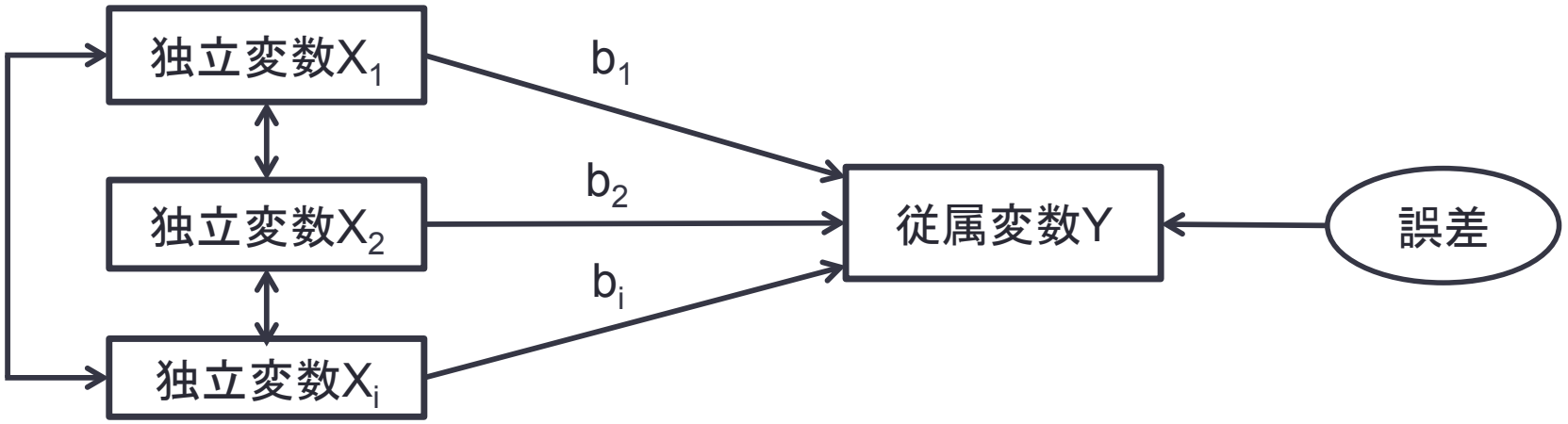
- 変数間の因果関係の方向性を仮定し、1つまたは複数の独立変数によって従属変数をどれくらい説明できるのかを検討する手法
  - 単回帰分析：独立変数が1つの場合
  - **重回帰分析**：独立変数が2つ以上の場合

（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



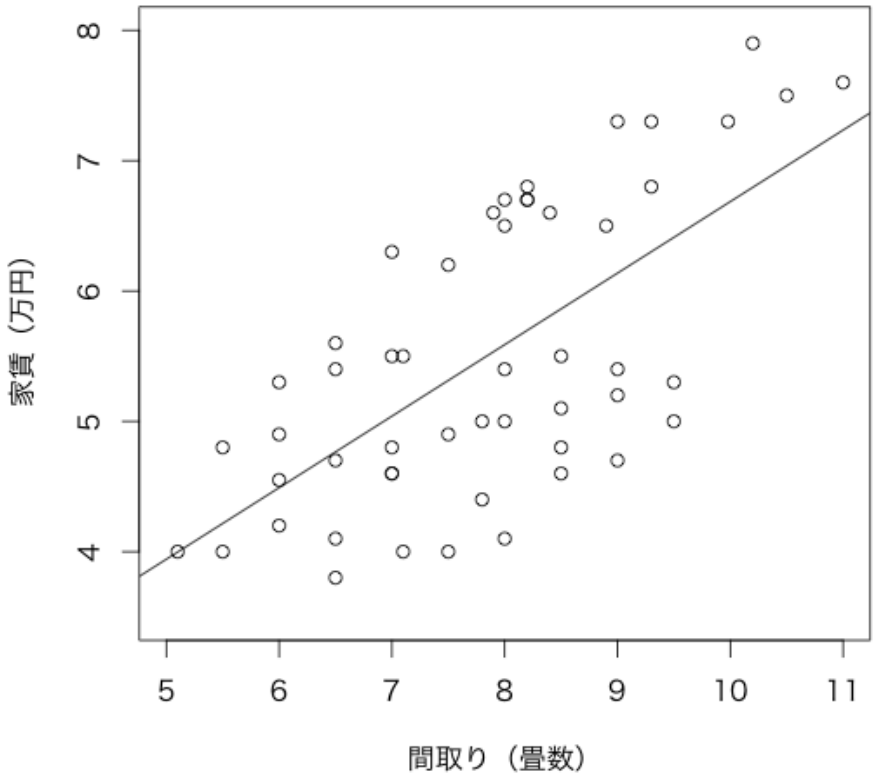
（復習）重回帰分析について

- 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）
  - 最小2乗法によって、残差（y-y^）が最も少なくなるようなa,bを求める
  - 回帰式の予測の精度は決定係数（R^2）



（復習）前回の発表での例

- 間取り（説明変数）から家賃（従属変数）を予測する場合
  - 単回帰式：$y=1.20+0.55x$
  - $R^2=0.43$ (分散の約43%を説明)

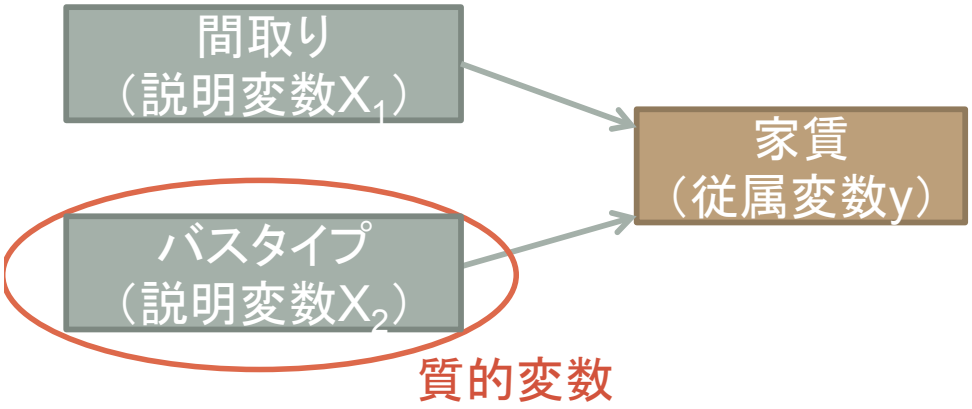


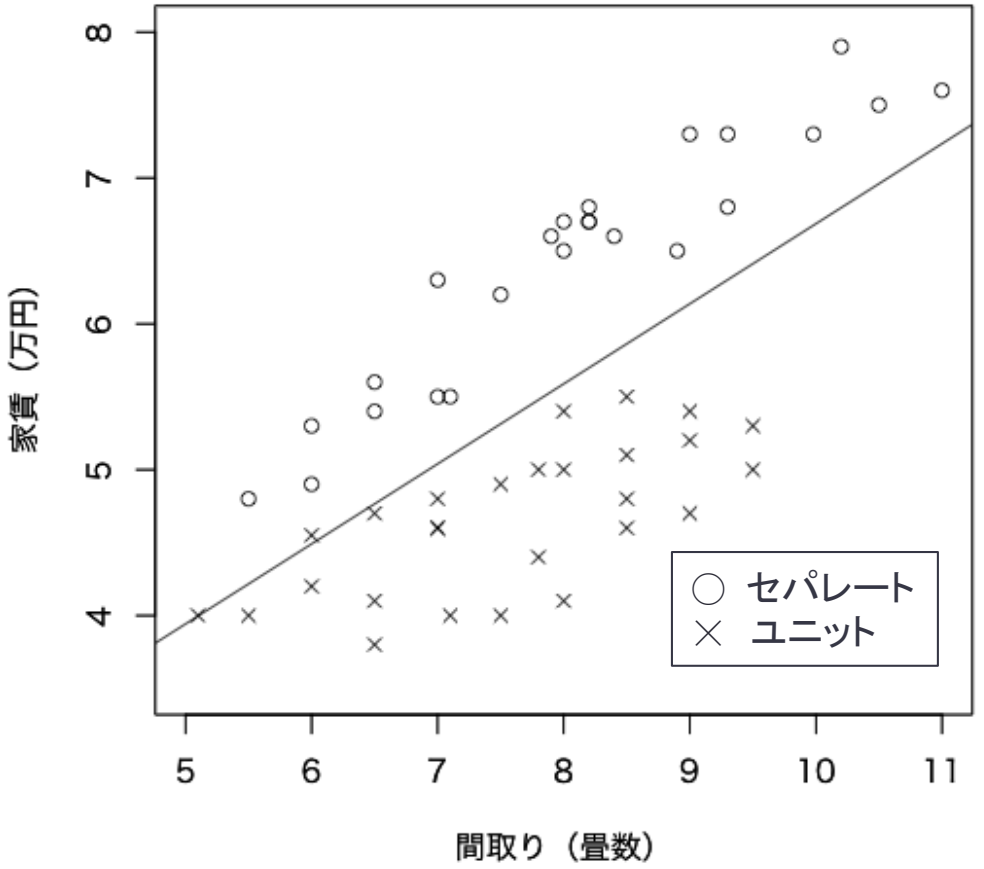
- 築年数も説明変数に加えると、、、
  - 重回帰式：$y=3.45+0.42x_1-0.06x_2$
  - 修正済み$R^2=0.64$

説明力（決定係数）が上昇！

（復習）カテゴリカル変数を含む回帰分析

- 説明変数にカテゴリカル変数が含まれる回帰分析





→ダミー変数を利用して、
変数の効果を検討する

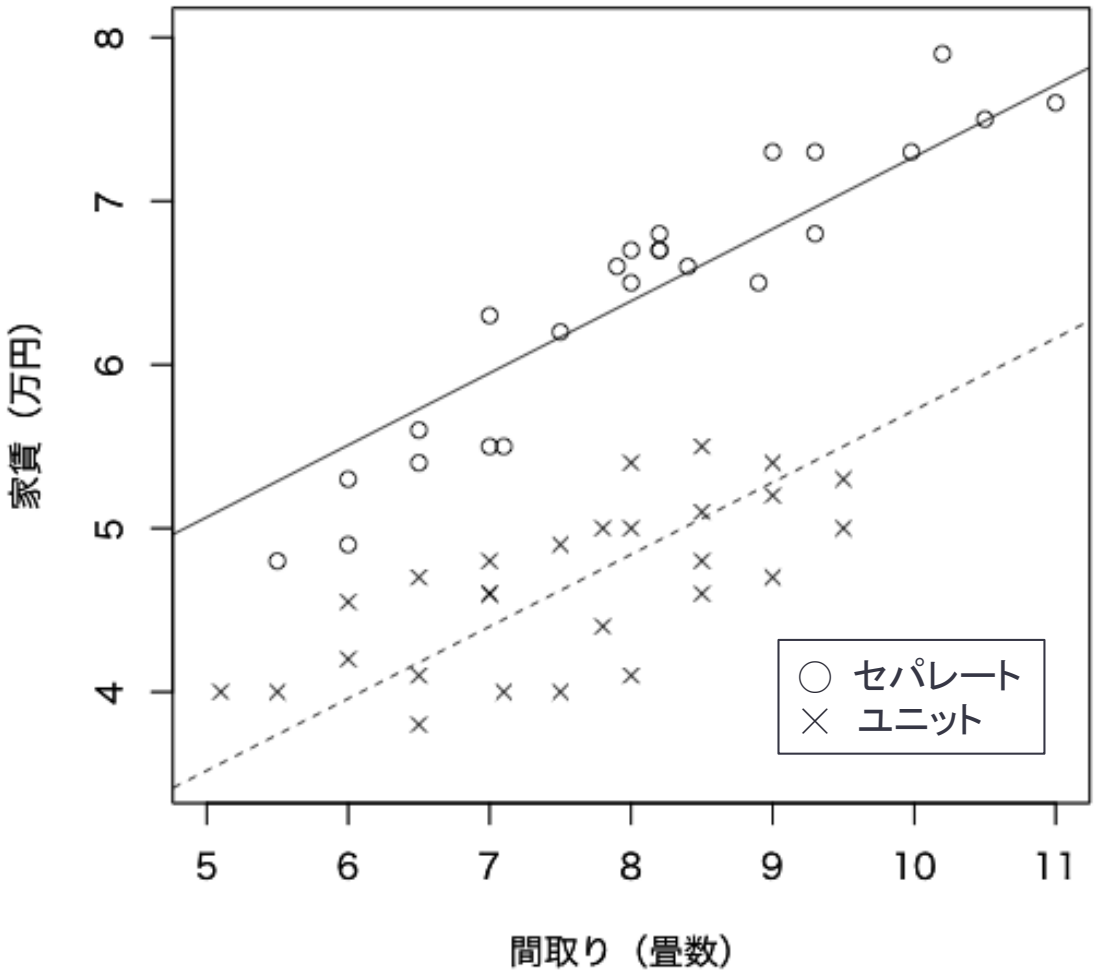
- $x_2 = \begin{cases} 0 & \text{セパレートバス} \\ 1 & \text{ユニットバス} \end{cases}$

（復習）カテゴリカル変数を含む回帰分析

- 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
  - $\hat{y} = a+b_1x_1+b_2x_2$
  - x_2=0の場合、
    - $\hat{y} = a+b_1x_1$
  - x_2=1の場合
    - $\hat{y} = a+b_2+b_1x_1$
    - と表される

- 重回帰式を求める
  - $y = 2.86 + 0.44x_1 - 1.55x_2$
    - x_2=0: $y = 4.61 + 0.26x_1$
    - x_2=1: $y = 2.71 + 0.26x_1$

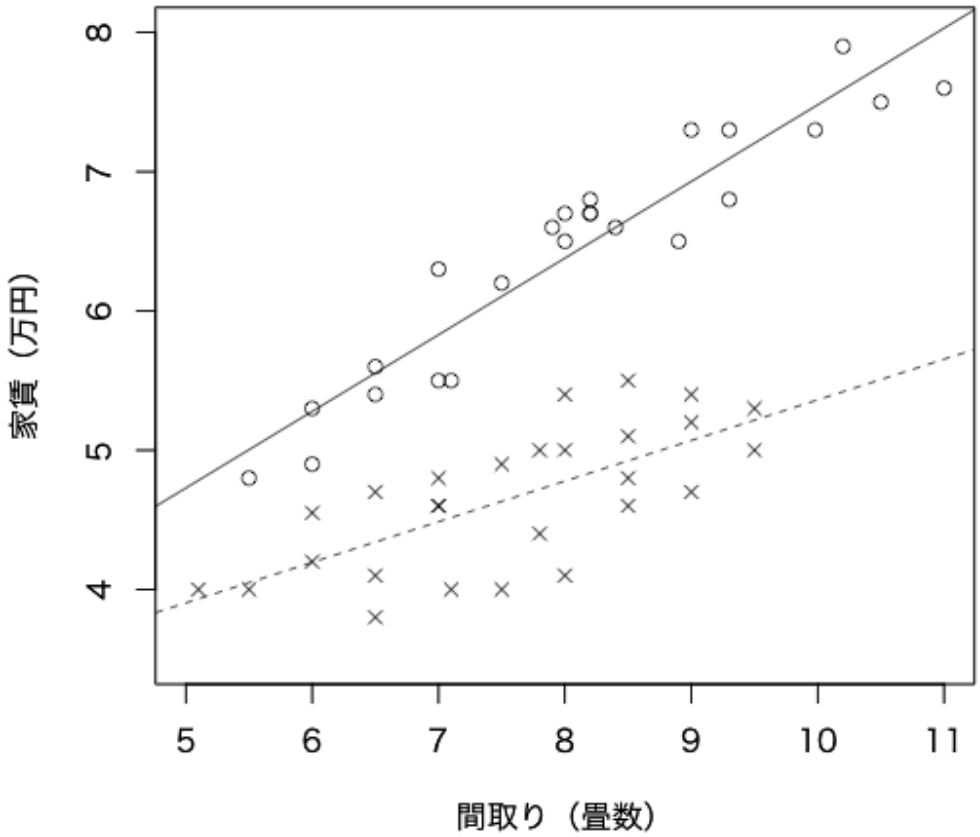
→決定係数：$R^{*2} = 0.89$



カテゴリカル変数を含む回帰分析

- カテゴリー間で**切片だけでなく傾き**も異なるモデルを想定できないか？

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



重回帰分析での交互作用の検討

- 一般的な重回帰分析では、各説明変数の主効果（偏回帰係数）がどの程度あるかを問題としており、交互作用は仮定されていない。

- 説明変数同士の積を新たな説明変数として組み込むことで、重回帰分析においても、交互作用を検討することが出来る

（その際、階層的重回帰分析によって、追加した交互作用項の有効性を検討する）

分析する上での注意点

- メリット（分散分析との違い）
  - 量的変数をそのまま予測変数として取り扱うことが出来る（グループ化する必要がない、検定力の低下を抑えられる）
  - 要因数が増えると多くのサンプルが必要になるが、サンプルサイズがそれほど多くなくても検定可能

- 問題点
  - 交互作用項の解釈の問題

    →むやみに入れない（そもそも偏回帰係数自体、解釈が難しい）
  - 交互作用項と説明変数間の相関が高くなってしまうため、多重共線性が生じやすくなる

    →中心化の必要性

交互作用項を含む重回帰モデル

カテゴリー間で切片および傾きが異なる重回帰モデルを、以下の式で表現する

- $\hat{y}=a+b_1x_1+b_2x_2+b_3x_1x_2$

この式をx_1でまとめると、x_1におけるyへの一次モデルとみなすことができ、

- $\hat{y}=(a+b_2x_2)+(b_1+b_3x_2)x_1$

x_1の係数にはb_3x_2が含まれるため、x_1の効果はx_2によって変化すること（＝交互作用効果）を示している

中心化（センタリング）

- 交互作用項（x_1x_2）はx_1とx_2で作られた変数であるため、x_1あるいはx_2 と相関が非常に高くなってしまう（→多重共線性の可能性）

- 交互作用項を含む重回帰分析を行うためには、あらかじめx_1とx_2に中心化（＝変数の平均値を0となるように変換する処理）を行う必要がある

- 主効果の項（x_1とx_2）と交互作用（x_1x_2）の項の相関による多重共線性を回避することが出来る
  - 中心化は「多重共線性を回避するテクニック」ではないことに注意
  - 数学的解説はAiken & West (1991)を参照

Rで分析

- データの読み込み
  - dat0<-read.csv(“dat.csv”)
  - y<-dat0$rent #従属変数
  - x1<-dat0$area1-mean(dat0$area1) #説明変数→中心化
  - x2<-dat0$bath #説明変数（ダミー変数）→中心化せず
- 中心化の確認
  - cor(data.frame(dat0$area1,dat0$bath,dat0$area1*dat0$bath))
  - cor(data.frame(x1,x2,x1*x2))

```
> cor(data.frame(area1,bath,area1*bath))
                  area1       bath area1...bath
area1         1.0000000 -0.1879078   -0.0416582
bath         -0.1879078  1.0000000    0.9752000
area1...bath -0.0416582  0.9752000    1.0000000
> cor(data.frame(x1,x2,x1*x2))
                 x1         x2    x1...x2
x1        1.0000000 -0.1879078  0.6594386
x2       -0.1879078  1.0000000 -0.1369229
x1...x2   0.6594386 -0.1369229  1.0000000
```

階層的重回帰分析

- 重回帰分析にはいくつかの変数選択の方法が存在
  - 総当り法や、逐次選択法（ステップワイズ法など）・・・

- 階層的重回帰分析
  - 予測変数の中で優先する変数による重回帰分析をまず行い、その後、興味の対象となる変数を加えた場合の決定変数R^2（平方和）の変化量によって、追加した変数の有効性を検討する

→今回、1次の項（主効果）であるx1とx2をまず組み入れ、その後、交互作用項であるx1x2を組み込んだ場合の決定係数（平方和）の変化量をもとに、交互作用項の有効性を検討する

Rで分析

- lm関数によるx1とx2の主効果のみの重回帰分析
  - reg1<-lm(y~x1+x2)
  - summary(reg1)

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74015 -0.30699  0.03857  0.27067  0.59094

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.31123    0.07675   82.23  < 2e-16 ***
x1           0.44055    0.03953   11.14 6.48e-15 ***
x2          -1.54900    0.10637  -14.56  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3724 on 48 degrees of freedom
Multiple R-squared:  0.8956, Adjusted R-squared:  0.8913
F-statistic: 205.9 on 2 and 48 DF,  p-value: < 2.2e-16
```

決定係数が上昇

Rで分析

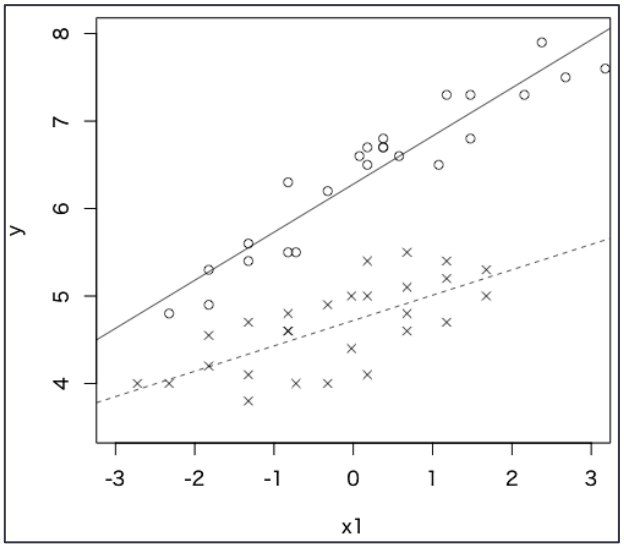
- lm関数による交互作用項を含む重回帰分析
  - reg2<-lm(y~x1*x2)
  - summary(reg2)

```
Call:
lm(formula = y ~ x1 * x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67840 -0.27464  0.06747  0.27164  0.62160

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.28187    0.06908  90.935  < 2e-16 ***
x1           0.55024    0.04656  11.817 1.12e-15 ***
x2          -1.55506    0.09509 -16.353  < 2e-16 ***
x1:x2       -0.25856    0.07149  -3.617 0.000726 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3329 on 47 degrees of freedom
Multiple R-squared:  0.9183, Adjusted R-squared:  0.9131
F-statistic: 176.2 on 3 and 47 DF,  p-value: < 2.2e-16
```



交互作用項の係数が有意！
（=傾きの差が有意）

決定係数は
さらに上昇

Rで分析

- 階層的重回帰分析によって、交互作用項の有効性を確認
- 決定係数（平方和）の変化量の検定
  - anova(reg1,reg2)

```
Analysis of Variance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 * x2
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     48 6.6571
2     47 5.2078  1    1.4493 13.08 0.0007265 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

決定係数は
有意に増加

交互作用の下位検定

- x_2がどのような値のときに、x_1の効果がどのようになるかを検討する

先ほどの結果より、求められた重回帰式は、

- $\hat{y} = 6.28+0.55x_1-1.56x_2-0.26x_1x_2$
- $\hat{y} = (6.28-1.56x_2) +(0.55-0.26x_2)x_1$

x_1のyへの回帰式とみなすことができ、x_1の係数である$(0.55-0.26x_2)$を**単純傾斜（simple slope）**という

交互作用の下位検定

- 特定のx_2の値を代入することで、それぞれのx_2の値でのx_1のyの回帰直線を求めることが出来る

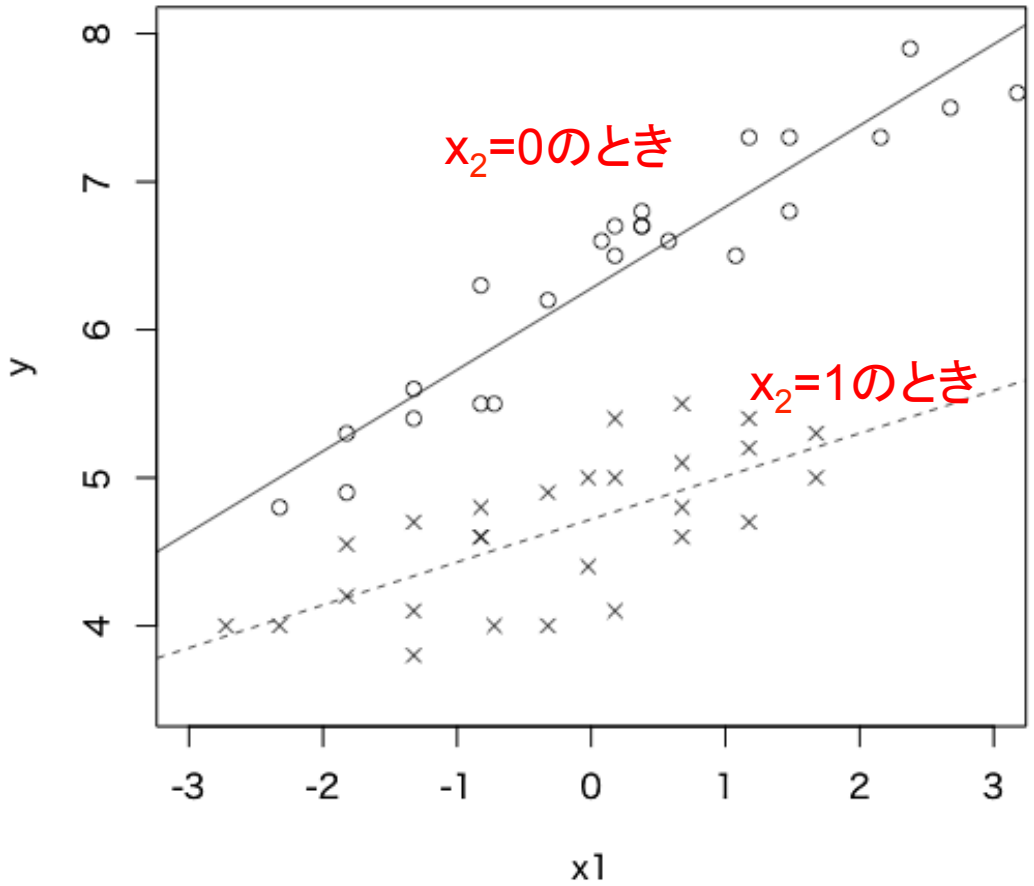
→今回はダミー変数の値（x_2=0, x_2=1）を代入

※平均値、平均+1SD、平均-1SDがCohen & Cohen(1983)によって提案されている

交互作用の下位検定

x_2にもとづくx_1のyへの回帰直線の算出

- x_2=0のとき: y=6.28+0.55x_1　#切片とx_1の係数
- x_2=1のとき: y=(6.28-1.56・1) +(0.55-0.26・1)x_1=4.72+0.29x_1



単純傾斜の有意性

- 求められた単純傾斜は、0.55, 0.29

→この単純傾斜は、統計的に有意な効果を持つのかどうかを検定する（単純傾斜の有意性の検定）

- このうち、x_2=0のときの単純傾斜である0.55は、x_1の主効果と同じであり、全体モデルの有意性検定においてすでに検定済み（ダミー変数を使用する利点）

単純傾斜の有意性

x_2=1のときの単純傾斜である0.29について検定

①まず、新たな変数zを作成し、

- $z=x_2-1$

とする（交互作用項をx_2=0, x_2=1で作成していたため）

②説明変数x_1と新たに作成したZを掛けあわせた変数x_1zを作成する

③x_1、z、x_1zを説明変数とする重回帰分析を行う

この分析結果において、x_2=1として算出したx_1のyへの単純傾斜（0.29）の有意性検定も行われる

Rで分析

- z=x2-1
- reg2<-lm(y~x1*z)
- summary(reg2)

```
Call:
lm(formula = y ~ x1 * z)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67840 -0.27464  0.06747  0.27164  0.62160

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72681    0.06535  72.332  < 2e-16 ***
x1           0.29168    0.05425   5.377 2.33e-06 ***
z           -1.55506    0.09509 -16.353  < 2e-16 ***
x1:z        -0.25856    0.07149  -3.617 0.000726 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3329 on 47 degrees of freedom
Multiple R-squared:  0.9183, Adjusted R-squared:  0.9131
F-statistic: 176.2 on 3 and 47 DF,  p-value: < 2.2e-16
```

x_2=1の単純傾斜の値と一致し、
有意性検定の結果も有意

終わりに

- 3値以上のカテゴリカル変数の分析について
  - 水準数-1のダミー変数を用いることで検討可能
  - （例）3群の場合: (d1,d2) = (0,0), (0,1), (1,0)
  - 交互作用の検討については、前田和寛さんのHPを参照（http://blog.kz-md.net/?p=711）

- 調整変数がカテゴリカル変数でない場合について
  - 調整変数についても、中心化を行う
  - 交互作用の下位検定を行う際に、「平均値」、「平均+1SD」、「平均-1SD」を代入（Cohen & Cohen(1983)により提案されている）

参考文献

- 豊田秀樹（2012）回帰分析入門-Rで学ぶ最新データ- 東京書籍
- 前田和寛（2008）重回帰分析の応用的手法--交互作用項ならびに統制変数を含む分析
- Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park, CA: Sage.
- Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.
- 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