2022年3月27日
第24回春の合宿セミナー（日本行動計量学会）
（統計的因果推論入門）

講義5
重回帰分析の限界

長崎大学 情報データ科学部 准教授
高橋 将宜
博士（理工学）
m-takahashi@nagasaki-u.ac.jp

概要

- 仮定1：誤差項の期待値はゼロ
- 仮定2：パラメータ（母数）における線形性
- 仮定3：誤差項の条件付き期待値ゼロ

教科書 Ch.7

- 仮定4：完全な多重共線性がない
- 仮定5：誤差項の分散均一性
- 仮定6：誤差項の正規性

教科書 Ch.8

- ここがポイント

仮定1：誤差項の期待値はゼロ

仮定の概要

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

- $E[\varepsilon_i] = 0$
  - 誤差項ε_iは期待値を取るとゼロ
- 誤差項ε_iは観測されない
  - これは紛れもなく仮定ではあるものの，この仮定を満たすことは難しくない.
- 誤差項の期待値がゼロでない場合
  - Y-切片の値がα_0から$\beta_0 = \alpha_0 + \delta$に変わる
  - 傾きはβ_1のまま
  - 影響を受けるのは，Y-切片の値だけであり，傾きβ_1には影響がなく，誤差項ε_iの期待値はゼロと見なすことができる

証明について

- 教科書pp.90-91

仮定2：パラメータ（母数）における線形性

仮定の意味と重要性

- 意味
  - 回帰モデルにおけるパラメータ（母数）が線形という仮定
  - 散布図で見たとき、XとYが直線的な関係に見えることが必要
- 重要性
  - 最小二乗法による回帰係数の推定量が不偏であるために必要

状況設定

- $Y_{1i} = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
- $Y_{2i} = \beta_0 + \beta_1 log(X_{2i}) + \varepsilon_i$
- $Y_{3i} = exp(\beta_0 + \beta_1 X_{1i} + \varepsilon_i)$
- $Y_{4i} = \beta_0 X_{2i}^{\beta_1} e^{\varepsilon_i}$

- $\beta_0 = 1.0$
- $\beta_1 = 1.5$
- $\varepsilon_i \sim N(0, 1)$
- $X_{1i} \sim N(0, 1)$
- $X_{2i} \sim LN(0, 1)$

X_{2i}は対数正規分布（lognormal distribution）に従っているとしよう．Xが対数正規分布に従っているとは，$log(X)$が正規分布に従っているという意味である（松原・縄田・中井1991, p.128）．

データの読み込み：data07a（教科書p.92）

- data07a <- read.csv(file.choose( ))
- attach(data07a)
- summary(data07a)

```
> summary(data07a)
       y1                y2                y3                 y4                  x1                  x2
 Min.   :-4.8549   Min.   :-4.3252   Min.   :   0.008   Min.   :  0.00487   Min.   :-3.25322   Min.   : 0.02902
 1st Qu.:-0.2441   1st Qu.:-0.2313   1st Qu.:   0.783   1st Qu.:  0.29191   1st Qu.:-0.68967   1st Qu.: 0.53061
 Median : 0.9851   Median : 1.0138   Median :   2.678   Median :  1.01390   Median :-0.03448   Median : 0.99452
 Mean   : 0.9640   Mean   : 1.0113   Mean   :  19.685   Mean   :  5.49271   Mean   :-0.01626   Mean   : 1.68969
 3rd Qu.: 2.2265   3rd Qu.: 2.3361   3rd Qu.:   9.267   3rd Qu.:  3.80417   3rd Qu.: 0.73734   3rd Qu.: 2.04450
 Max.   : 8.2378   Max.   : 6.7444   Max.   :3781.174   Max.   :312.44944   Max.   : 3.63957   Max.   :17.49898
```

$$Y_{1i} = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$
$$Y_{2i} = \beta_0 + \beta_1 log(X_{2i}) + \varepsilon_i$$
$$Y_{3i} = exp(\beta_0 + \beta_1 X_{1i} + \varepsilon_i)$$
$$Y_{4i} = \beta_0 X_{2i}^{\beta_1} e^{\varepsilon_i}$$

散布図（1）

$Y_{1i} = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

図1

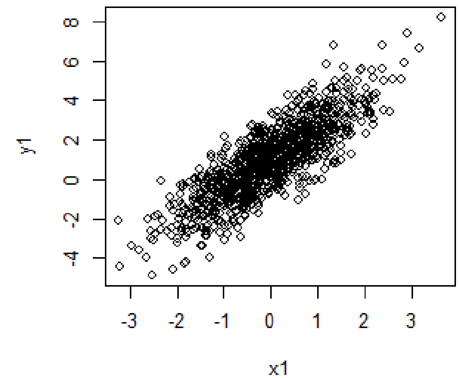
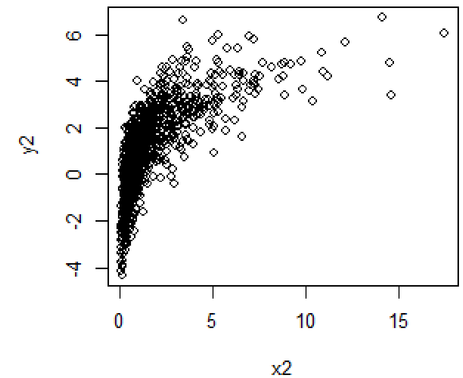


図2

$Y_{2i} = \beta_0 + \beta_1 X_{2i} + \varepsilon_i$



$X_{2i} \sim LN(0, 1)$

$Y_{3i} = exp(\beta_0 + \beta_1 X_{1i} + \varepsilon_i)$

図3

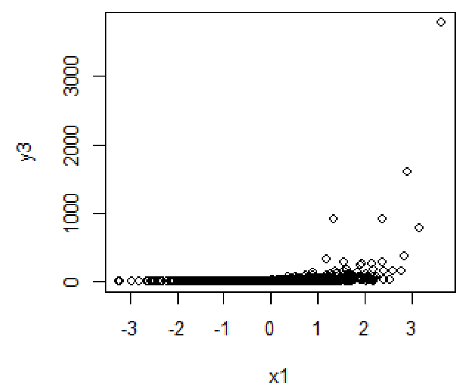
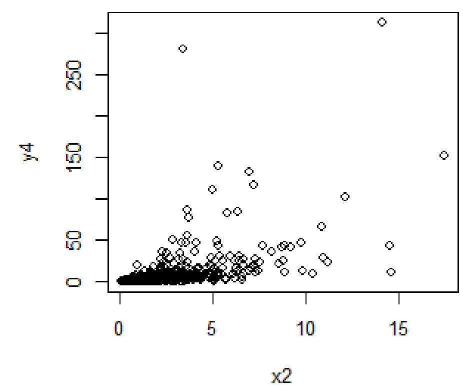


図4

$Y_{4i} = \beta_0 X_{2i}^{\beta_1} e^{\varepsilon_i}$



Rコード：散布図（1）

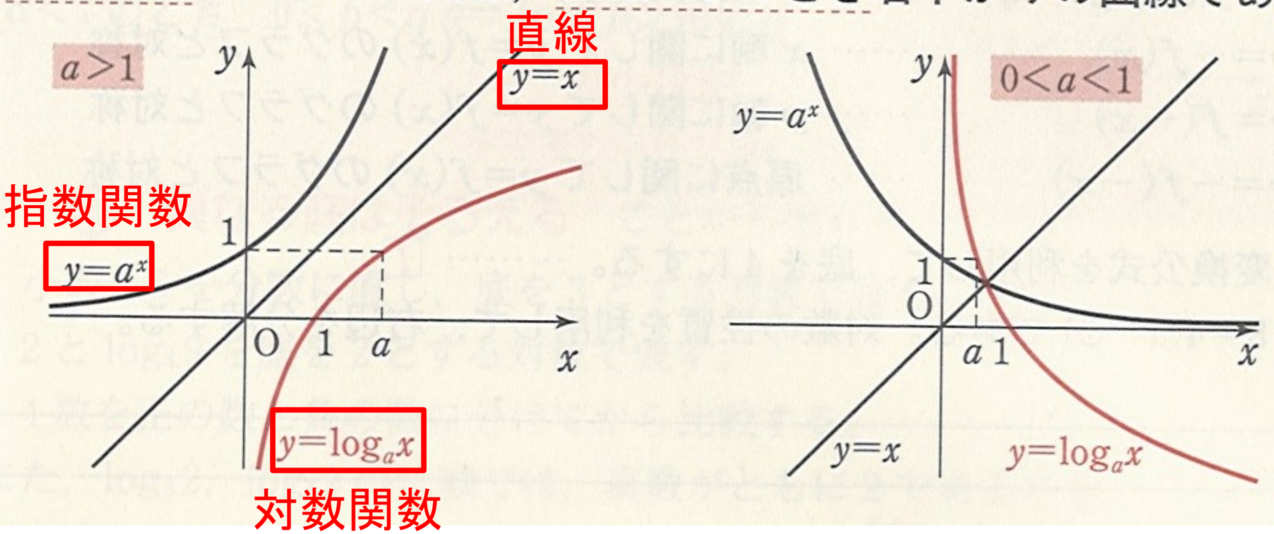
```
layout(matrix(1:4, 2, 2, byrow=TRUE))
plot(x1, y1, main="図1")
plot(x2, y2, main="図2")
plot(x1, y3, main="図3")
plot(x2, y4, main="図4")
```

高校数学の復習：指数関数と対数関数のグラフ

対数関数のグラフ　対数関数 $y=\log_a x$ のグラフ は，指数関数 $y=a^x$ のグラフと，直線 $y=x$ に関して対称で，次のようになる。

1　点 $(1,\ 0)$，$(a,\ 1)$ を通り，y 軸を漸近線とする曲線である。

2　$a>1$ のとき右上がりの曲線，$0<a<1$ のとき右下がりの曲線である。



出典：『チャート式　基礎からの数学II+B』（数研出版, 2017）

仮定2：パラメータ（母数）における線形性

散布図と関数形

$Y_{1i} = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

図1

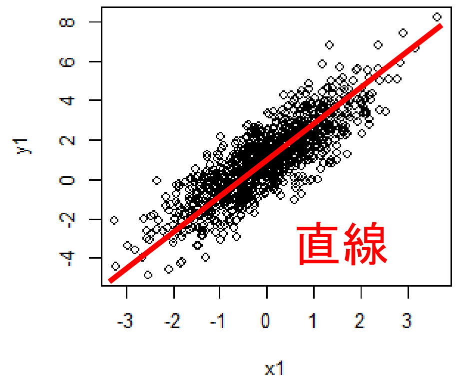
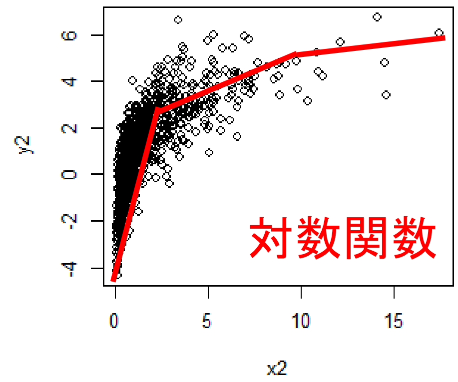


図2

$Y_{2i} = \beta_0 + \beta_1 X_{2i} + \varepsilon_i$



$X_{2i} \sim LN(0,1)$

$Y_{3i} = exp(\beta_0 + \beta_1 X_{1i} + \varepsilon_i)$

図3

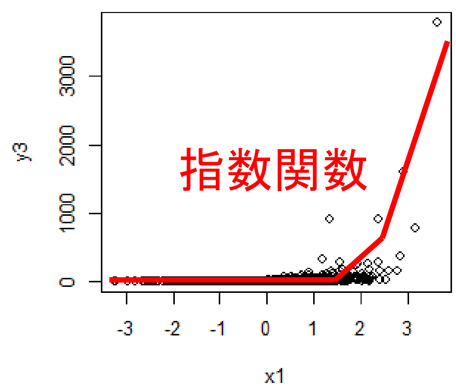
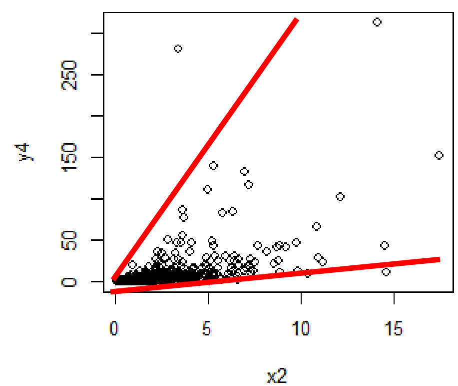


図4

$Y_{4i} = \beta_0 X_{2i}^{\beta_1} e^{\varepsilon_i}$



不均一分散とも関連あり

データ解析例（1a）（教科書p.93）

$\beta_0 = 1.0,\ \beta_1 = 1.5$

- model1 <- lm(y1 ~ x1)
- model2 <- lm(y2 ~ x2)
- model3 <- lm(y3 ~ x1)
- model4 <- lm(y4 ~ x2)

```
summary(model1)
summary(model2)
summary(model3)
summary(model4)
```

	モデル1	モデル2	モデル3	モデル4
Y-切片	0.988	-0.105	20.277	-3.352
傾き	1.506	0.661	36.446	5.235

Model 1

XとYは線形の関係にあるため，モデル1では，$\hat{\beta}_1 = 1.506$となっており，$\beta_1 = 1.5$を正しく推定できている様子が分かる．Y-切片も0.988なので，$\beta_0 = 1.0$を正しく推定できている．

Model 2～Model 4

XとYは非線形の関係にあった．

したがって，それぞれ，$\hat{\beta}_1 = 0.661$，$\hat{\beta}_1 = 36.446$，$\hat{\beta}_1 = 5.235$となっており，$\beta_1 = 1.5$をうまく推定できていない様子が分かる．

データ解析例（1b）（教科書p.93）

$\beta_0 = 1.0,\ \beta_1 = 1.5$

- model5 <- lm(y2 ~ log(x2))
- model6 <- lm(log(y3) ~ x1)
- model7 <- lm(log(y4)~log(x2))

```
summary(model5)
summary(model6)
summary(model7)
```

	モデル5	モデル6	モデル7
Y-切片	0.988	0.988	-0.012
傾き	1.550	1.506	1.550

1行目では，x2をlog関数で自然対数に変換している．
2行目では，y3を自然対数に変換している．
3行目ではy4とx2の両方を自然対数に変換している．
モデル5～モデル7では，それぞれ，$\hat{\beta}_1 = 1.550$，$\hat{\beta}_1 = 1.506$，$\hat{\beta}_1 = 1.550$となっており，$\beta_1 = 1.5$を正しく推定できている様子が分かる．細かな違いは標本抽出誤差である．

仮定2：パラメータ（母数）における線形性

散布図（2）

$$Y_{1i} = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

図1

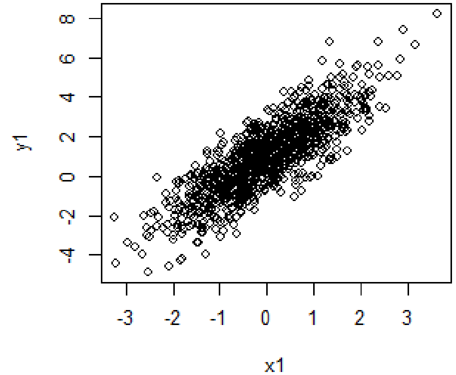
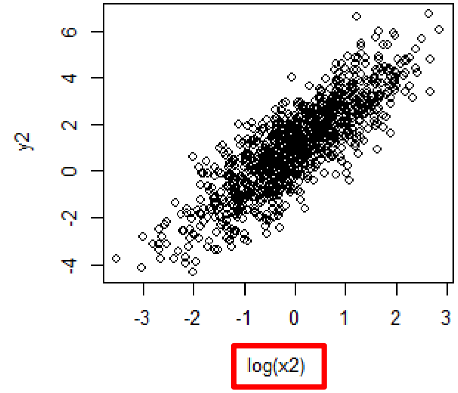


図2b

$$Y_{2i} = \beta_0 + \beta_1 X_{2i} + \varepsilon_i$$



$$X_{2i} \sim LN(0, 1)$$

$$Y_{3i} = exp(\beta_0 + \beta_1 X_{1i} + \varepsilon_i)$$

図3b

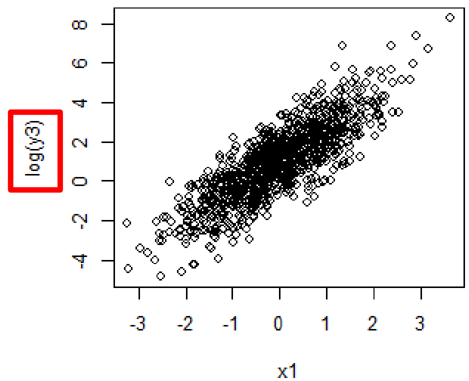
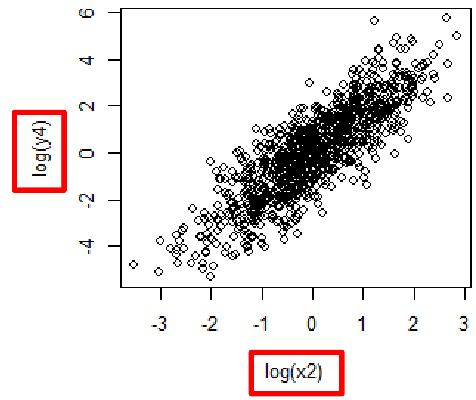


図4b

$$Y_{4i} = \beta_0 X_{2i}^{\beta_1} e^{\varepsilon_i}$$



Rコード：散布図（2）

- layout(matrix(1:4, 2, 2, byrow=TRUE))
- plot(x1, y1, main="図1")
- plot(log(x2), y2, main="図2b")
- plot(x1, log(y3), main="図3b")
- plot(log(x2), log(y4), main="図4b")

ここがポイント

- 重回帰モデルにおいて適切な因果推論を行うには，散布図で見たとき，XとYが直線的な関係に見えるような関数形を選択する必要がある．
  - そのために，変数を適切に変換する必要がある．

多変量の場合の診断方法

- ここまで，簡単のため二変量に限定して議論してきた．
- 単回帰モデルの場合，XとYの散布図から視覚的に関数の形を把握することができる．
- しかし，共変量が多変量である重回帰分析では，他の共変量を統制した場合の効果（偏回帰係数）に興味があるため，二変量の散布図では，適切な関数の形を探すことができない．
- そこで，成分プラス残差プロット（component-plus-residual plot）を使用することが推奨されている．

ほぼ実行不可能

- 重回帰モデルから平均処置効果（ATE）を推定するには，関数形を正しく設定することが重要である．
- ただし，成分プラス残差プロットを使用するには，組み合わせを解析者が考えて実行する必要がある．
- ゆえに，変数が増えてくると組み合わせを考えてモデルを組み上げる手間が膨大なものとなる．
- さらに，今回は関数形の候補として変換なしと対数変換の2つのみを考慮したが，考え得る関数形の候補は無数にある．

仮定3：誤差項の条件付き期待値ゼロ

仮定の内容

- $E[\varepsilon_i|\boldsymbol{X}] = E[\varepsilon_i]$
  - 誤差項ε_iと共変量$\boldsymbol{X}$は独立という仮定
  - この仮定が満たされているとき，誤差項ε_iは共変量$\boldsymbol{X}$と平均独立（mean independent）という

- $E[\varepsilon_i|\boldsymbol{X}] = 0$
  - 仮定1（誤差項の期待値ゼロ）より，$E[\varepsilon_i] = 0$
  - 共変量$\boldsymbol{X}$が与えられたとき，誤差項ε_iの期待値はゼロ
  - このとき，誤差項ε_iは共変量$\boldsymbol{X}$と条件付き平均独立（conditional mean independent）という

仮定の意味と重要性

- 最小二乗法による回帰係数の推定量が不偏であるために必要な仮定
  - 統計的因果推論において非常に重要
- 説明変数X_1と共変量X_2に相関がある場合
  - X_2をモデルに含めないと，X_1からYへの効果には交絡が起こっていた
  - X_2をモデルに含めることで，重回帰モデルでは，X_1からYへの純粋な効果を推定できた
- 仮定3（誤差項の条件付き期待値ゼロ）の意味
  - 交絡因子が十分にモデルに含まれていなければならない

診断方法

- 診断方法はない
- 誤差項ε_i
  - 観測されないため，直接的には検証できない
- 残差e_iと共変量$\boldsymbol{X}$との関係
  - 残差と説明変数は，最小二乗法の原理によってそもそも相関がないので，自動的に$E[e_i|\boldsymbol{X}]=0$となってしまう

無視可能な割付け

- $Pr(T_i|Y_i(1), Y_i(0), \boldsymbol{X}) = Pr(T_i|\boldsymbol{X})$

- 観測された共変量の値が同じ個体同士では，処置の割付けは無作為化されていると考えてよいという仮定
  - 共変量$\boldsymbol{X}$に十分な変数が含まれているかどうかが重要

多数の共変量

- この仮定を満たすには？
  - できるだけ多くの変数をモデルに取り入れて，必要な共変量を取りこぼす可能性を下げる必要がある
- 疑問点
  - データセット内にある変数は，すべてモデルに取り込んでしまってもいいのか？
- 検討すべきこと
  - 不要な変数をモデルに取り入れることの影響
  - 因果関係の間に位置する変数の取り扱い

不要な変数をモデルに取り入れる問題

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$
  - $\beta_0 = 1.0$
  - $\beta_1 = 1.3$
  - $\beta_2 = 1.2$
  - $\beta_3 = 0.0$
  - $\varepsilon_i \sim N(0, 1)$
- 興味の対象：X_1からYへの因果効果β_1
- $\beta_3 = 0.0$
  - X_3は不要な変数
  - 不要な変数X_3をモデルに取り込んだ場合，β_1の推定にどのような影響が出るか？

結論

- β_1の推定は偏りなく行うことができる
  - X_1とX_3の相関が強くなると，標準誤差が大きくなるおそれがある.
- ある変数が不要と分かっているならば，モデルに入れる必要がないことは明らか
  - しかし，実証研究では，ある共変量が実際に必要なのか不必要なのか，分からないことが多い
  - 不要な変数をモデルに取り入れることには，不偏性という点からは大きな問題はないが，標準誤差に悪影響が出るおそれはある

データの読み込み：data07c（教科書p.101）

- data07c <- read.csv(file.choose( ))
- attach(data07c)
- summary(data07c)

```
> summary(data07c)
       y1                  x1                   x2                   x3
 Min.   :-8.1670   Min.   :-3.013259   Min.   :-2.795193   Min.   :-3.39119
 1st Qu.:-0.6056   1st Qu.:-0.719596   1st Qu.:-0.700067   1st Qu.:-0.68897
 Median : 1.0544   Median :-0.009873   Median : 0.016571   Median :-0.04079
 Mean   : 1.0114   Mean   :-0.008575   Mean   : 0.004854   Mean   :-0.02090
 3rd Qu.: 2.7284   3rd Qu.: 0.703909   3rd Qu.: 0.707734   3rd Qu.: 0.62961
 Max.   : 8.5708   Max.   : 4.043700   Max.   : 3.606328   Max.   : 3.65575
```

$\beta_0 = 1.0$, $\beta_1 = 1.3$, $\beta_2 = 1.2$, $\beta_3 = 0.0$

データ解析例（2）（教科書p.102）

- model8 <- lm(y1 ~ x1)
- model9 <- lm(y1 ~ x1 + x2)
- model10 <- lm(y1 ~ x1 + x2 + x3)

```
summary(model8)
summary(model9)
summary(model10)
```

	モデル8	モデル9	モデル10
β_0	1.028	1.017	1.017
β_1	1.903 (0.045)	1.334 (0.036)	1.333 (0.074)
β_2		1.183	1.183
β_3			0.000

- $\beta_1 = 1.3$であるから，モデル8の推定結果1.903は誤り
- モデル9の推定結果1.334とモデル10の推定結果1.333は，正しい
- 不要な変数x3をモデルに含めても，偏りは発生していない

問題点

- 不要な変数x3をモデルに含めることに悪影響はないのだろうか？
- confint関数を用いて，モデル9とモデル10の95%信頼区間を計算
  - モデル8については，そもそも点推定値が誤っているので，ここでは信頼区間の検討はしていない．
- confint(model9)
- confint(model10)

信頼区間の幅

$\beta_0 = 1.0,\ \beta_1 = 1.3,\ \beta_2 = 1.2,\ \beta_3 = 0.0$

```
> confint(model9)
                2.5 %   97.5 %
(Intercept) 0.9525962 1.081587
x1          1.2629626 1.404155
x2          1.1117615 1.254657
>
> confint(model10)
                 2.5 %    97.5 %
(Intercept)  0.9525525 1.0816372
x1           1.1877819 1.4787193
x2           1.0952147 1.2714541
x3          -0.1294489 0.1300779
```

- モデル9における$\hat{\beta}_1$の95%信頼区間は1.263～1.404
- モデル10における$\hat{\beta}_1$の95%信頼区間は1.188～1.479
- どちらも真値$\beta_1 = 1.3$を区間の中に含んでいることから，正しい結果ではあるが，モデル10の信頼区間の方が，モデル9の信頼区間よりも幅が大きい
- 推定の精度に影響があり
- 不要な変数x3をモデルに取り込んだ結果，推定結果にノイズが混ざったため，標準誤差に悪影響が出た

標準誤差の大きさ

$\beta_0 = 1.0, \ \beta_1 = 1.3, \ \beta_2 = 1.2, \ \beta_3 = 0.0$

- model8 <- lm(y1 ~ x1)
- model9 <- lm(y1 ~ x1 + x2)
- model10 <- lm(y1 ~ x1 + x2 + x3)

```
summary(model8)
summary(model9)
summary(model10)
```

	モデル8	モデル9	モデル10
β_0	1.028	1.017	1.017
β_1	1.903 (0.045)	1.334 (0.036)	1.333 (0.074)
β_2		1.183	1.183
β_3			0.000

- モデル9の標準誤差は0.036
- モデル10の標準誤差は0.074

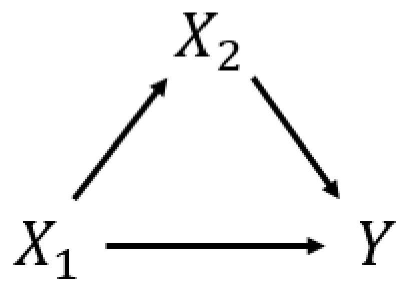
結論を再確認

- 不要な変数をモデルに取り入れても偏りはないため，ある変数が必要かどうか判断に迷う場合は，入れる方がよい
- ただし，その変数が実際には不要だった場合，偏りには問題がないものの，推定の精度に影響が出ることには注意が必要であり，信頼区間が必要以上に大きくなっているおそれはある

中間変数をモデルに取り入れる問題

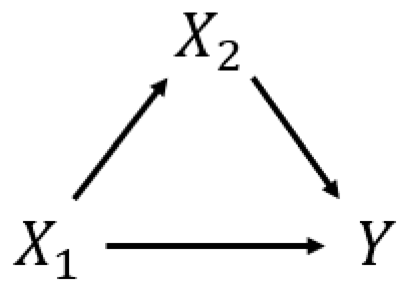
- 中間変数
  - 因果関係の間に位置する変数
- 結論
  - 中間変数はモデルに含めてはならない

方向付き非巡回グラフ（DAG）



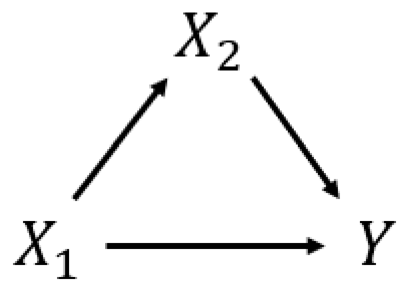
- Xが原因で，Yが結果であるとき，$X \to Y$と表すもの
  - X_1がYの原因と考えている
- 以前のDAG：$X_1 \leftarrow X_2$
- 今回のDAG：$X_1 \to X_2$

中間変数



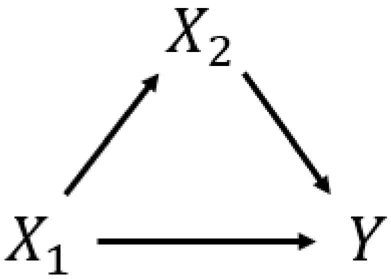
- X_2はX_1とYの間にある
- X_2のような因果関係の間に位置する変数を中間変数（mediator）といい，共変量とは別のものとして考える

全体の効果＝直接効果+間接効果



- 直接効果（direct effect）
  - $X_1 \rightarrow Y$の部分
- 間接効果（indirect effect）
  - $X_1 \rightarrow X_2 \rightarrow Y$の部分
- 全体の効果（total effect）
  - 直接効果と間接効果を合わせたもの

中間変数のあるDAGを式で表す



- $X_{2i} = \gamma_0 + \gamma_1 X_{1i} + \varepsilon_{1i}$
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_{2i}$

全体の効果

- X_1からYへの全体の効果は$\beta_1 + \beta_2\gamma_1$である.
- これが推定対象である.

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_{2i}$
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2(\gamma_0 + \gamma_1 X_{1i} + \varepsilon_{1i}) + \varepsilon_{2i}$
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2\gamma_0 + \beta_2\gamma_1 X_{1i} + \beta_2\varepsilon_{1i} + \varepsilon_{2i}$
- $Y_i = (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)X_{1i} + (\beta_2\varepsilon_{1i} + \varepsilon_{2i})$

データの読み込み：data07d（教科書p.105）

- data07d <- read.csv(file.choose( ))
- attach(data07d)
- summary(data07d)

```
> summary(data07d)
       y1                x1                  x2
 Min.   :-8.8865   Min.   :-3.00805   Min.   :-4.8032
 1st Qu.:-0.2536   1st Qu.:-0.69737   1st Qu.:-0.2237
 Median : 2.2271   Median :-0.03532   Median : 0.9982
 Mean   : 2.1597   Mean   :-0.01165   Mean   : 0.9663
 3rd Qu.: 4.5914   3rd Qu.: 0.68843   3rd Qu.: 2.1525
 Max.   :16.8370   Max.   : 3.81028   Max.   : 8.0559
```

- x2=1.0+1.5*x1+e1として生成
  - x2はx1の影響を受けており，その影響は$\gamma_1 = 1.5$
- y1=1.0+1.3*x1+1.2*x2+e2として生成
  - x1はy1に対して直接効果$\beta_1 = 1.3$を与えている
  - x2もy1に対して間接効果$\beta_2 = 1.2$を与えている
- 推定対象は$\beta_1 + \beta_2\gamma_1$である
  - $1.3 + 1.2 \times 1.5 = 3.1$

データ解析例（3）

$\boxed{\beta_1 + \beta_2\gamma_1 = 3.1}$

- model11 <- lm(y1 ~ x1)
- model12 <- lm(y1 ~ x1 + x2)

```
summary(model11)
summary(model12)
```

	モデル11	モデル12
β_0	2.196	0.994
β_1	3.157	1.316
β_2		1.222

- 推定すべき因果効果の真値は3.1であった
- モデル11では，3.157であるから，正しく推定できている
- モデル12では，1.316であり，正しく推定できていない
  - ◆ ここで注意すべきことは，1.316はβ_1の値として正しく，これは$X_1 \to Y$の直接効果を表している
- しかし，推定すべき因果効果は，直接効果と間接効果を合わせた全体の効果であり，これは$\beta_1 + \beta_2\gamma_1 = 3.1$の方だった

結論を再確認

- 中間変数をモデルに含めてしまうと，推定対象である全体の効果を適切に推定できなくなってしまうので，中間変数はモデルに入れてはならない

統制すべき共変量に関するまとめ

- 教科書pp.131-133

復習：講義4

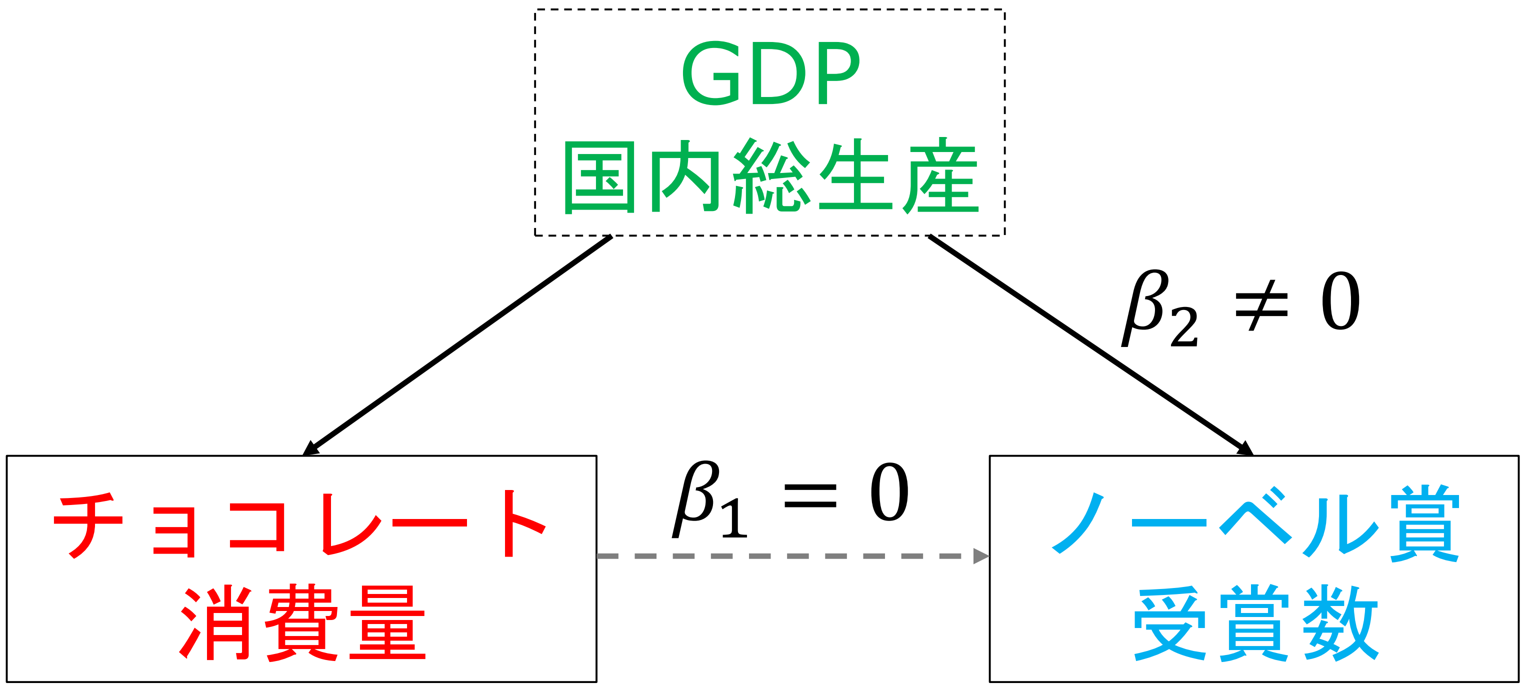
仮定3：誤差項の条件付き期待値ゼロ

重回帰モデル

共変量の解釈を行わない理由

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

$$\widehat{\text{ノーベル賞受賞数}}_i = \beta_0 + \beta_1 \text{チョコレート消費量}_i + \beta_2 GDP_i$$

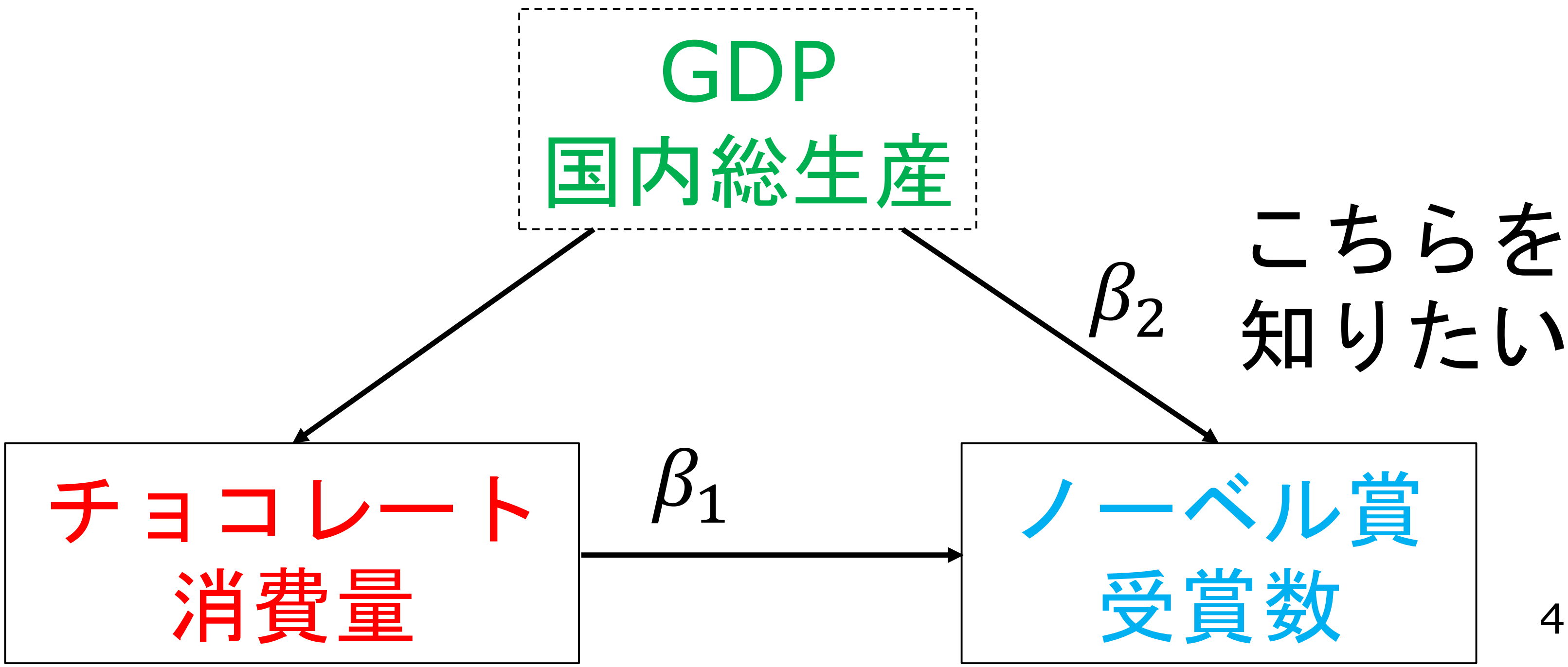


中間変数の問題

共変量の解釈を行わない理由

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

$$\widehat{\text{ノーベル賞受賞数}}_i = \beta_0 + \beta_1 \text{チョコレート消費量}_i + \beta_2 GDP_i$$



復習：講義4　　仮定3：誤差項の条件付き期待値ゼロ

重回帰モデル（教科書p.87）

共変量の解釈を行わない理由

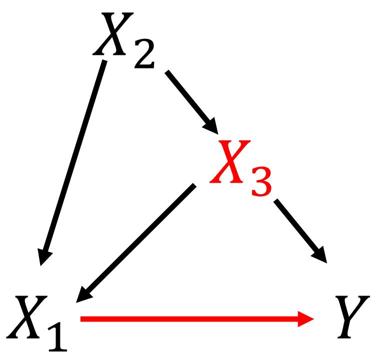
- model3<-lm(y1~x1+x2)
- summary(model3)
- confint(model3, level=0.95)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.32035    3.20331  -1.973   0.0625 .
x1           1.50477    0.75619   1.990   0.0604 .
x2           0.19552    0.08531   2.292   0.0329 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.064 on 20 degrees of freedom
Multiple R-squared:  0.5985,    Adjusted R-squared:  0.5583
F-statistic:  14.9 on 2 and 20 DF,  p-value: 0.0001089

> confint(model3, level=0.95)
                   2.5 %    97.5 %
(Intercept) -13.00233992 0.3616436
x1           -0.07260349 3.0821496
x2            0.01757188 0.3734626
```

仮定3：誤差項の条件付き期待値ゼロ

複雑な中間変数1

このケースについては，林・黒木（2017, p.36, pp.40-41, IWANAMI DATA SCIENCE Vol.3）も参照されたい.

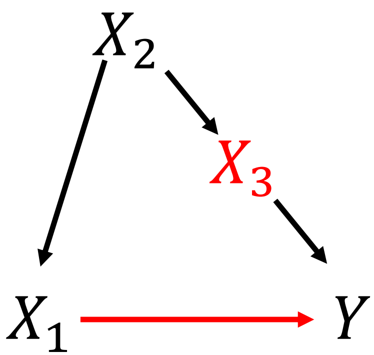



```
set.seed(1)

n1<-1000

x2<-rnorm(n1)

e1<-rnorm(n1)
e2<-rnorm(n1)
e3<-rnorm(n1)

x3<-x2+e1
x1<-x2+x3+e2

y1<-x1+x3+e3

summary(model1<-lm(y1~x1+x2+x3))
summary(model2<-lm(y1~x1+x2))
summary(model3<-lm(y1~x1+x3))
```

	Estimate	Std. Error
model1	1.009	0.032
model2	1.517	0.027
model3	1.009	0.026

x1からy1への真の因果効果は1.0

複雑な中間変数2

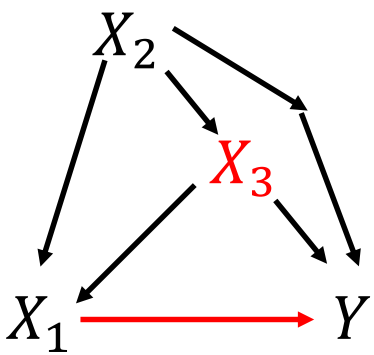



```
set.seed(1)

n1<-1000

x2<-rnorm(n1)

e1<-rnorm(n1)
e2<-rnorm(n1)
e3<-rnorm(n1)

x3<-x2+e1
x1<-x2+e2

y1<-x1+x3+e3

summary(model1<-lm(y1~x1+x2+x3))
summary(model2<-lm(y1~x1+x2))
summary(model3<-lm(y1~x1+x3))
```

	Estimate	Std. Error
model1	1.009	0.032
model2	1.032	0.046
model3	1.009	0.026

x1からy1への真の因果効果は1.0

複雑な中間変数3

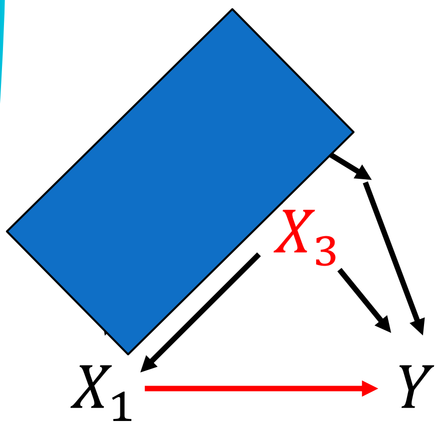



```
set.seed(1)

n1<-1000

x2<-rnorm(n1)

e1<-rnorm(n1)
e2<-rnorm(n1)
e3<-rnorm(n1)

x3<-x2+e1
x1<-x2+x3+e2

y1<-x1+x2+x3+e3

summary(model1<-lm(y1~x1+x2+x3))
summary(model2<-lm(y1~x1+x2))
summary(model3<-lm(y1~x1+x3))
```

	Estimate	Std. Error
model1	1.009	0.032
model2	1.517	0.027
model3	1.347	0.030

x1からy1への真の因果効果は1.0

複雑な中間変数3の続き（1）



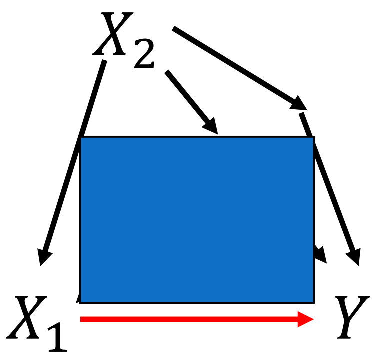

```
set.seed(1)

n1<-1000

x2<-rnorm(n1)

e1<-rnorm(n1)
e2<-rnorm(n1)
e3<-rnorm(n1)

x3<-x2+e1
x1<-x2+x3+e2

y1<-x1+x2+x3+e3
```

x1からy1への真の因果効果は1.0

```
summary(model1<-lm(y1~x1+x2+x3))
summary(model2<-lm(y1~x1+x2))
summary(model3<-lm(y1~x1+x3))
```

	Estimate	Std. Error
model1	1.009	0.032
model2	1.517	0.027
model3	1.347	0.030

複雑な中間変数3の続き（2）




```
set.seed(1)

n1<-1000

x2<-rnorm(n1)

e1<-rnorm(n1)
e2<-rnorm(n1)
e3<-rnorm(n1)

x3<-x2+e1
x1<-x2+x3+e2

y1<-x1+x2+x3+e3

summary(model1<-lm(y1~x1+x2+x3))
summary(model2<-lm(y1~x1+x2))
summary(model3<-lm(y1~x1+x3))
```

	Estimate	Std. Error
model1	1.009	0.032
model2	1.517	0.027
model3	1.347	0.030

x1からy1への真の因果効果は1.0

仮定4：完全な多重共線性がないこと

完全な多重共線性

- 2つ以上の説明変数に完全な相関がある

$$r_{x_1,x_2} = 1.0$$

- この場合、$\hat{\beta}$を推定することはできない
- 最小二乗法の仮定を満たしていない

多重共線性

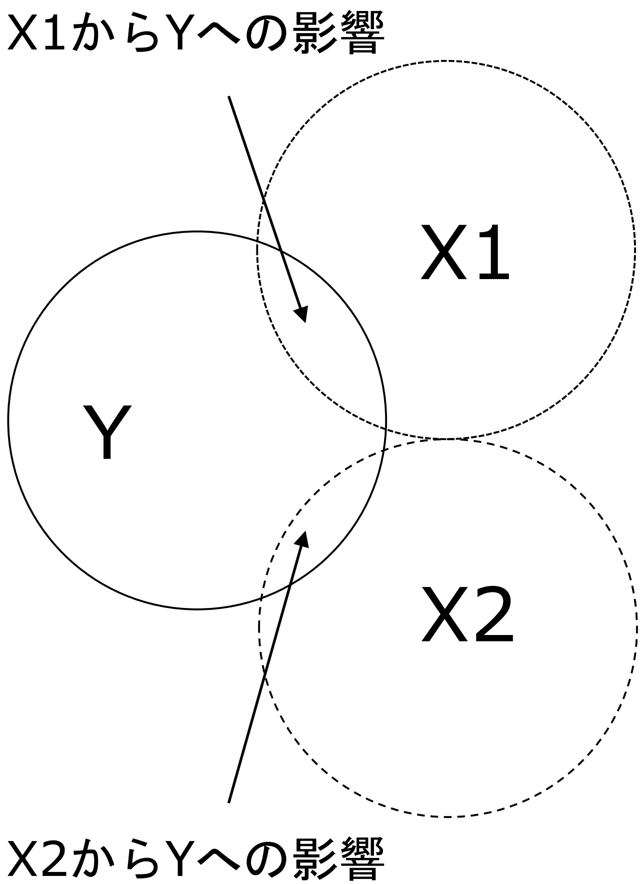
- 2つ以上の説明変数の相関が非常に強い

$$r_{x_1,x_2} \approx 1.0$$
$$r_{x_1,x_2} \neq 1.0$$

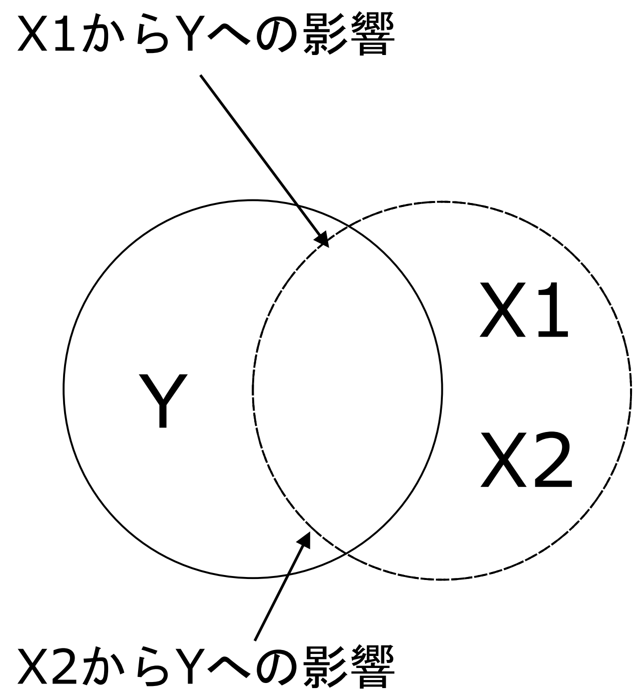
- この場合、$\hat{\beta}$を推定することはできる
- 最小二乗法の仮定は満たされている
- ただし、x_1とx_2のそれぞれの影響力を分離することが難しい

多重共線性を図示1

多重共線性なし

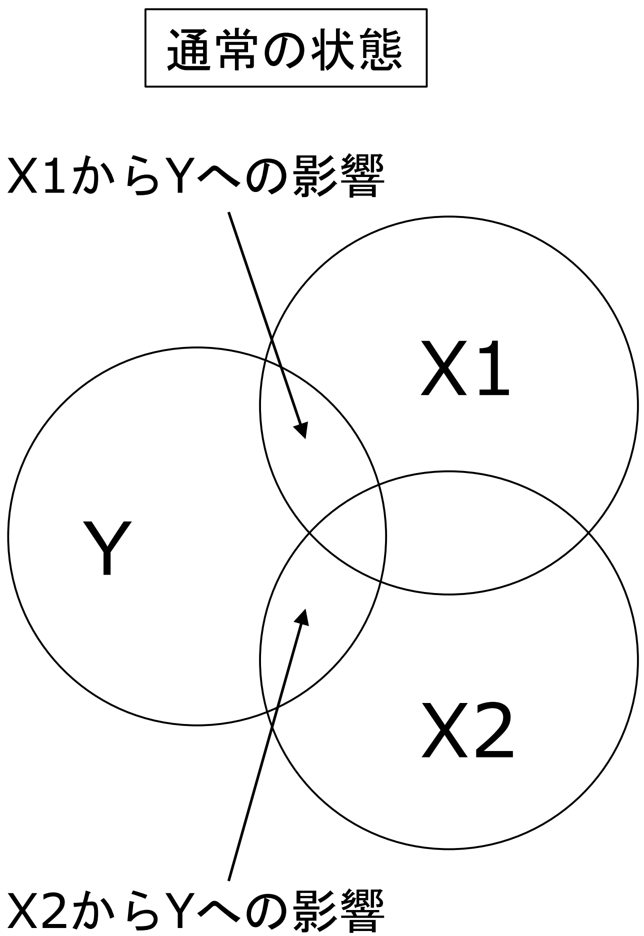


完全な多重共線性

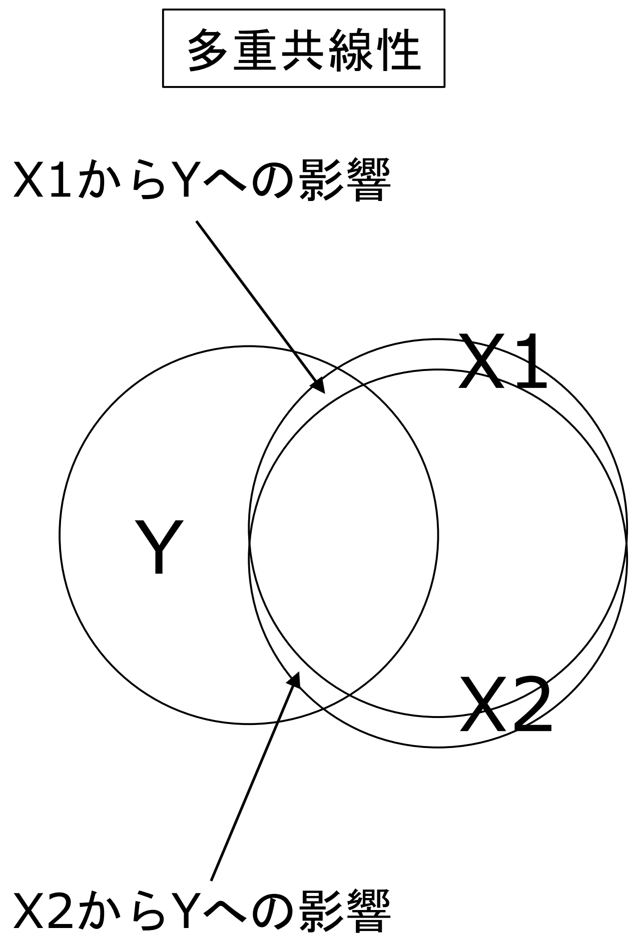


X1とX2の円は完全に重なっている。

多重共線性を図示2



通常の状態
X1からYへの影響
X1
Y
X2
X2からYへの影響



多重共線性
X1からYへの影響
X1
Y
X2
X2からYへの影響

多重共線性の影響

- よい結果
  - 回帰係数は不偏
- 悪い結果
  - X1からYへの影響とX2からYへの影響を区別することが難しい
  - X1からYへの影響が、実際に母集団においてあったとしても、帰無仮説を棄却できない可能性が高くなる。
  - 帰無仮説が正しくなかったとしても、信頼区間の中に0が含まれる可能性が高くなる。

診断方法と対処法

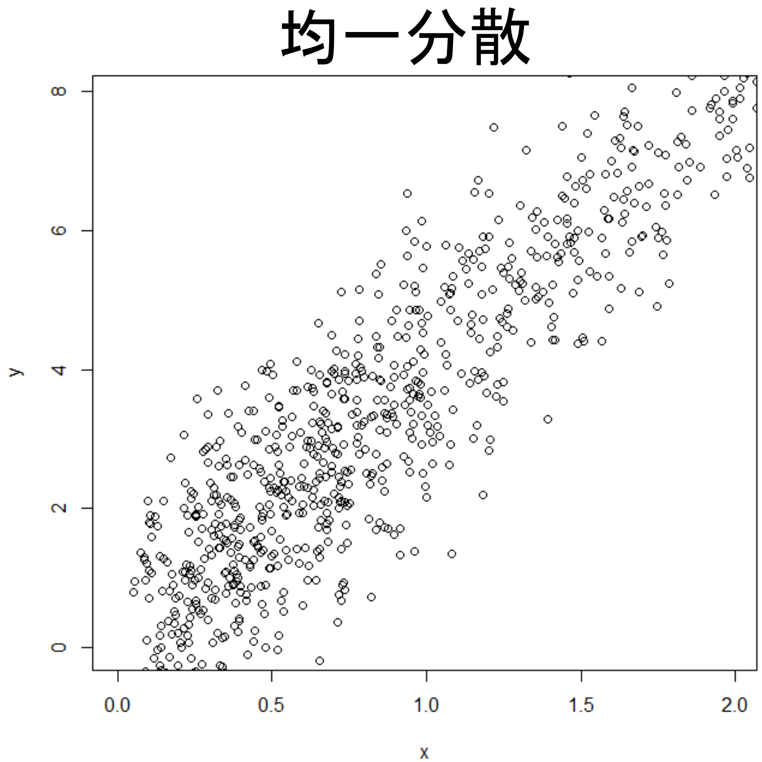
- 教科書pp.107-112

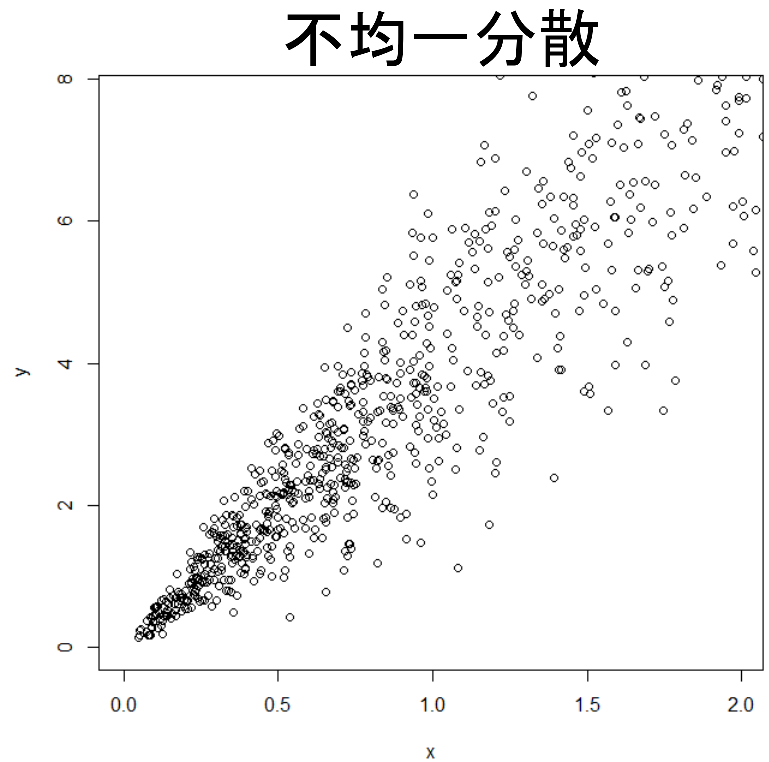
仮定5：誤差項の分散均一性

均一分散と不均一分散

- 均一分散（Homoskedasticity）
  - 説明変数の値にかかわらず、誤差項ε_iの分散は一定
  - $var(\varepsilon_i|X_i) = \sigma^2$
- 不均一分散（Heteroskedasticity）
  - 説明変数の値が変化すると、誤差項ε_iの分散も変化する
  - $var(\varepsilon_i|X_i) = \sigma_i^2$

不均一分散を図示1



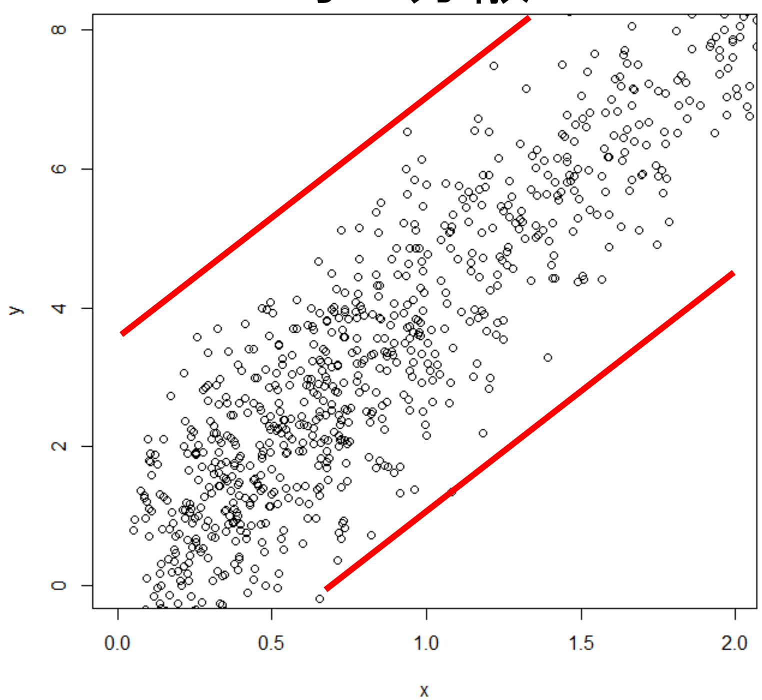


Xが変化しても、Yの分散は変化しない。

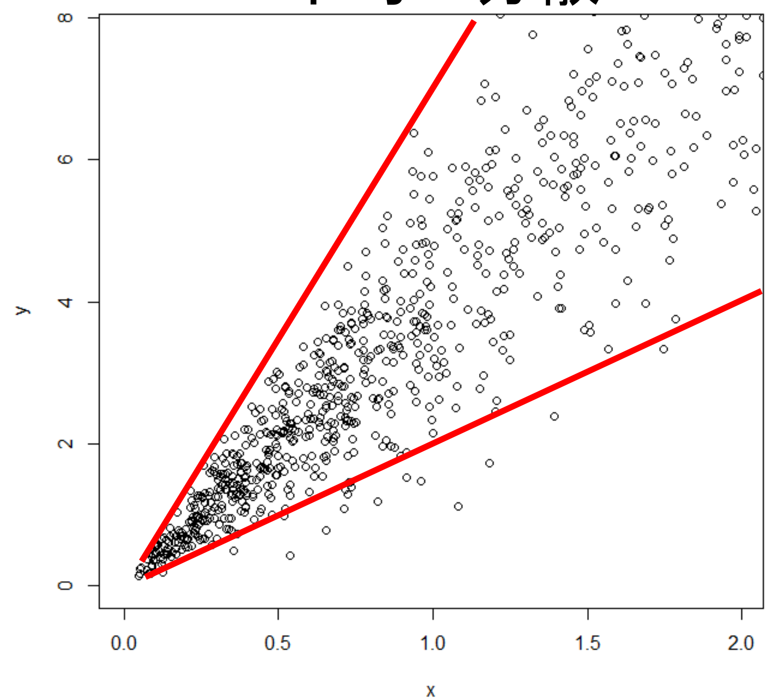
Xが変化すると、Yの分散も変化する。

不均一分散を図示2

均一分散



不均一分散

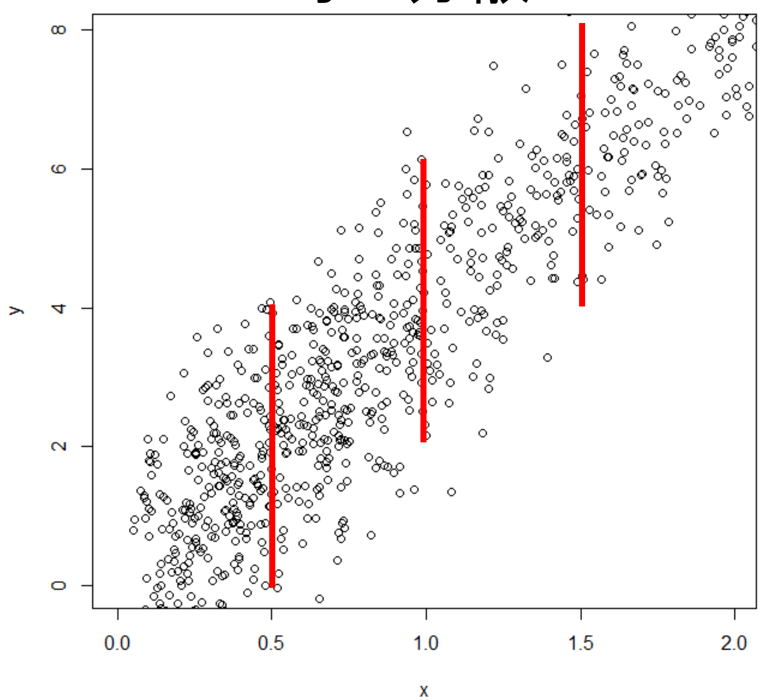


赤線はほぼ平行

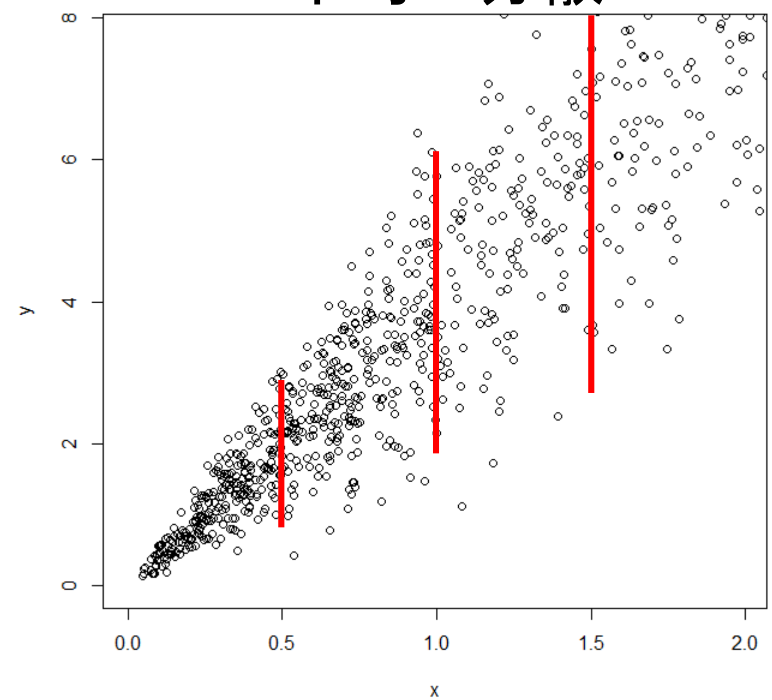
赤線は並行ではない

不均一分散を図示3

均一分散



不均一分散



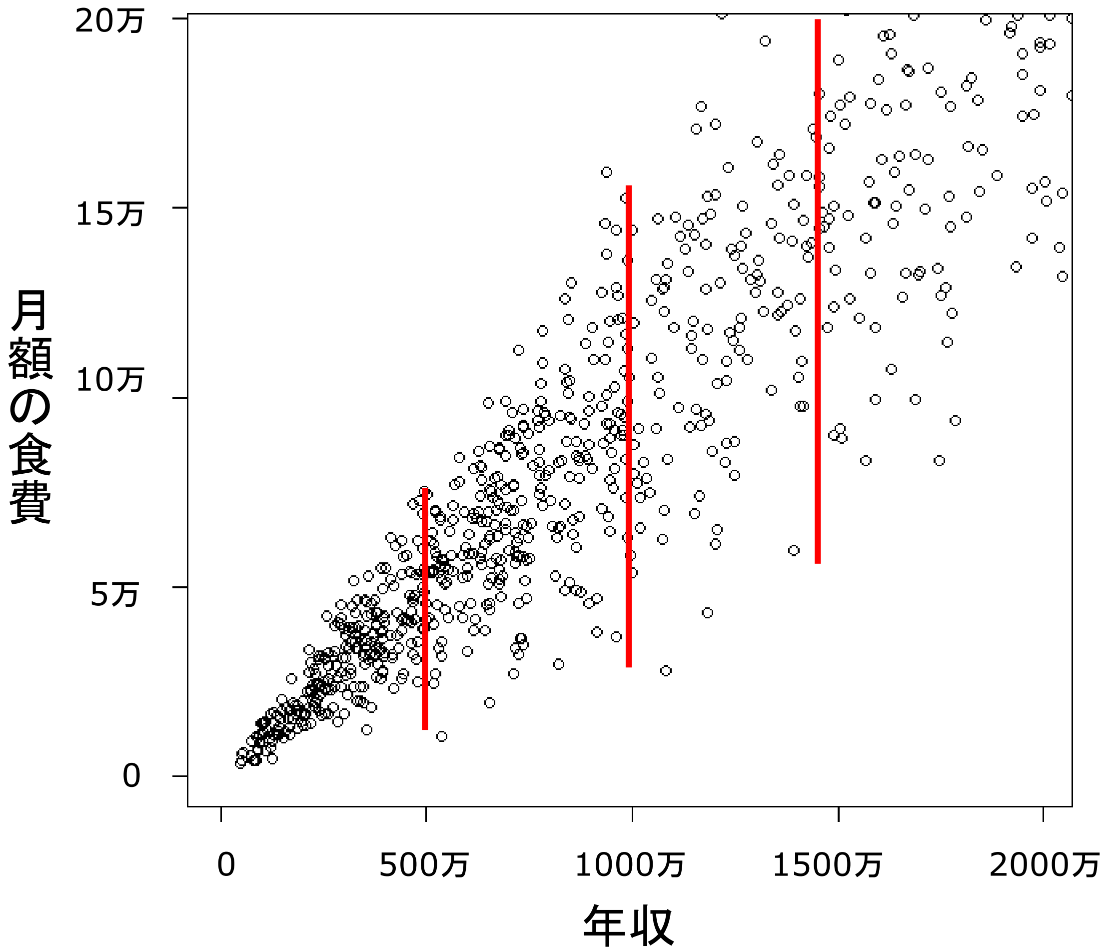
Xが大きくなっても、赤い縦線の長さは一定

Xが大きくなると、赤い縦線の長さも変化する。

不均一分散とは、結局、何？

- 不均一分散は、説明変数の値に依存して、より多くの裁量があることを示唆している。
- 例：収入と食事
  - 収入が少ない人：米やパンといった基本的な食事にしかお金を使うことができない。
  - 収入が多い人：いろいろな種類の食事にお金を使うことができる。
    - あるときにはファーストフードを食べたり、あるときには高級フレンチを食べたりできる。
    - 食事に費やす金額は、収入に依存してばらつきが大きくなる。

収入と食費



不均一分散の影響

- よい結果
  - 回帰係数は不偏
- 悪い結果
  - 標準誤差は不正確
  - 信頼区間も誤り
  - 仮説検定の結果は信頼できない

診断方法と対処法

- 教科書pp.112-120

仮定6：誤差項の正規性

追加の仮定

- 誤差項の正規性は，最小二乗法による推定量がBLUE（Best Linear Unbiased Estimator：最良線形不偏推定量）であるためには必須のものではない
- しかし，小標本において信頼区間の信頼度を名目どおりにするために必要な仮定である

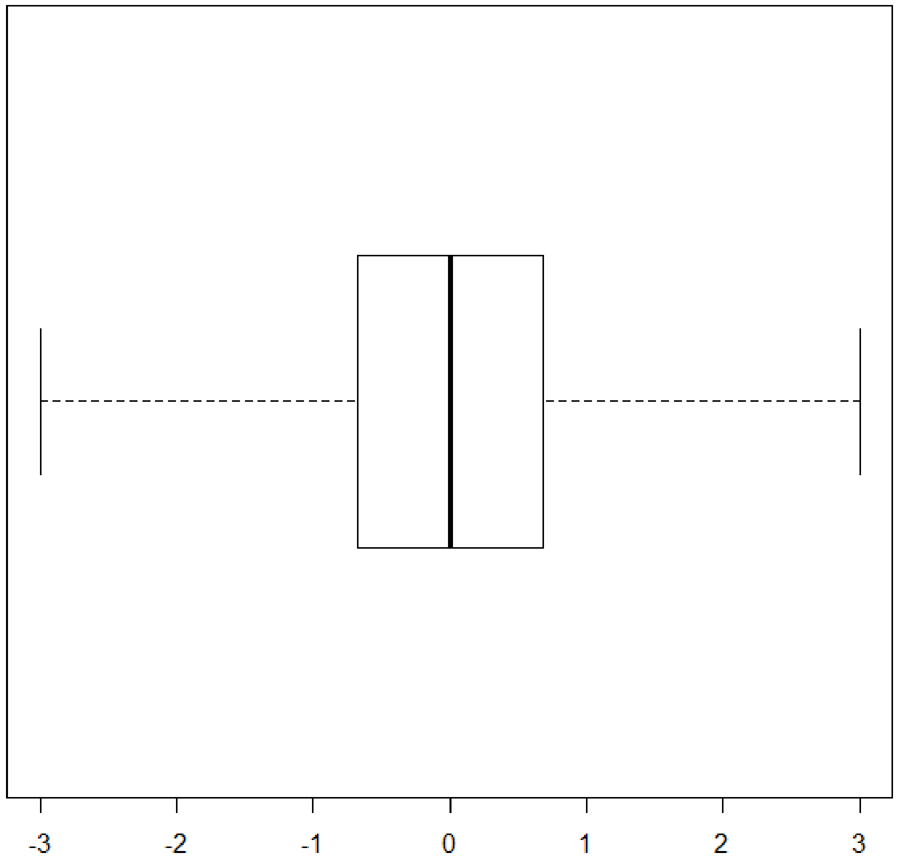
簡易な診断

- 回帰分析の出力結果に表示されている残差（Residuals）の基本統計量をチェックする

```
> summary(model1)

Call:
lm(formula = price ~ distance + roomsize)

Residuals:
   Min     1Q Median     3Q    Max
 -4736  -2472     25   1243   5172

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1417.7     3154.4   0.449  0.65847
distance     -1961.6      924.1  -2.123  0.04790 *
roomsize       761.2      177.8   4.281  0.00045 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3072 on 18 degrees of freedom
Multiple R-squared:  0.5854,	Adjusted R-squared:  0.5393
F-statistic: 12.71 on 2 and 18 DF,  p-value: 0.0003623
```

誤差項が正規分布している場合

```
e<-c(-3.00, -0.68, -0.00, 0.68, 3.00)
boxplot(e, horizontal=TRUE, range=0)
```



- 箱が均一に広がっている。
- 残差eは正規分布している。
- 誤差項εも正規分布している可能性が高い。

注：range=0とは、箱ひげ図に外れ値を表示しないようにするオプション

診断方法と対処法

- 教科書pp.120-123

ここがポイント

共分散分析の限界

- 仮定2を満たすためには，さまざまな関数形を試して，成分プラス残差プロットで確認する必要があった．
- 仮定3を満たすために，中間変数や操作変数に注意を向けながら，共変量をできるだけ多くモデルに取り入れる必要があった．
  - 共変量Xは多変量であるから，膨大な組み合わせのモデリングを考慮しなければならない
  - よって，共変量Xが多変量のときにも対応できるような，フレキシブルな方法が望ましい
  - 傾向スコアモデリングは，共分散分析と比較して，モデルの誤設定に対して強い

重回帰モデルの重要性

- ただし，ここまで扱ってきた重回帰モデルは，統計的因果推論の役に立たないという意味ではない．
- 傾向スコアモデリングにおける解析モデルは，回帰モデルの形をとっている．
- 操作変数法の推定方法として二段階最小二乗法という手法を用いるが，これは本質的に重回帰モデルの拡張である．
- 回帰不連続デザインは，名前のとおり，重回帰モデルの拡張である．