回帰分析のモデル検査

データ解析演習 2011/6/22

M2 後藤崇志

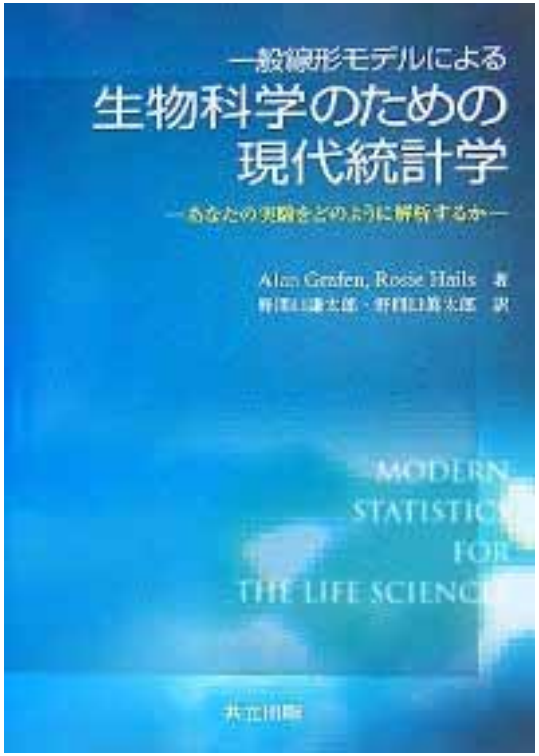
回帰分析とは・・・

- （複数の）独立変数から、従属変数を予測する統計手法
- “従属変数＝傾き1*独立変数1+傾2*独2＋・・・＋切片”のようなモデル式を求める
- モデルは、個々のデータからの誤差が最小になるように求められる
- 今日はそんな回帰分析の、あまり陽に当たらない部分についての話です

本日の内容

- 重回帰分析で満たすべき仮定について説明
- ↑の仮定が満たされているか検査する方法を説明
- 仮定が満たされていない場合の解決策を提案
- 実際にやってみる

- なお、内容は
  『一般線形モデルによる
  生物化学のための現代統計学』
  の7章、9章の一部をまとめた
  ものが中心です



重回帰分析で満たされるべき仮定

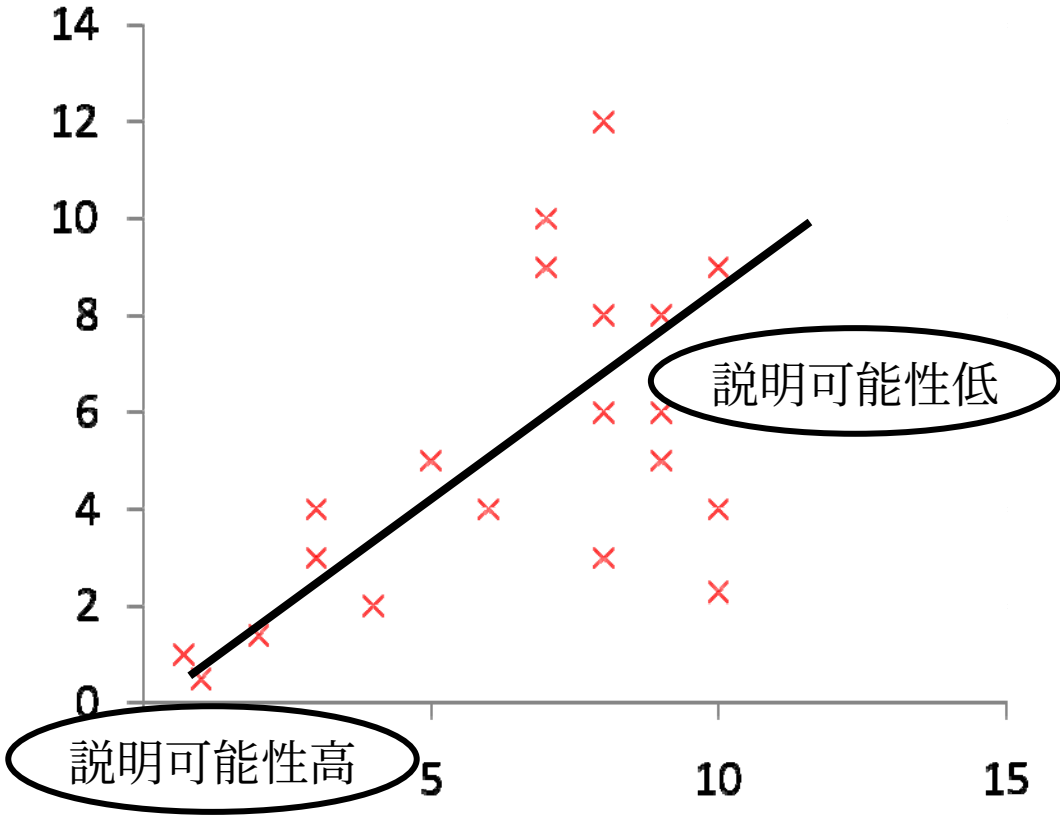
- 独立性
  - 「部分集合がどのように選ばれても、その部分集合がもつ誤差についての知識が他のデータ点の誤差について何の情報も与えない」
    → 今回は扱わない

- 分散の均一性
- 誤差の正規性
- モデルの線形性
  ① なぜ満たされていないと駄目なのか？
  ② 満たされているかどうかはどうやって判断するのか？
  ③ 満たされていない場合はどうしたらいいのか？

- **今回は数式ではなく、主として図を用いて考えてみる**

分散の均一性

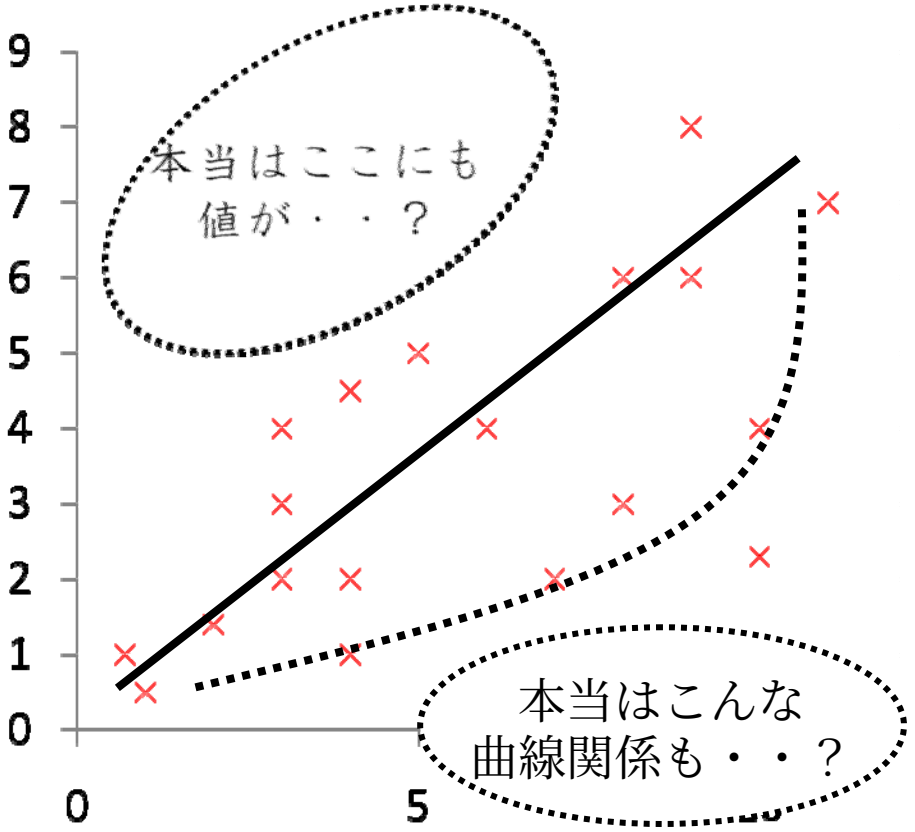
- 独立変数と従属変数の間に関連があるならば、独立変数により説明される変動＞誤差により説明される変動となる
  - ただし、分散が均一であるという仮定のもとで成り立つ
  - 分散が不均一な場合は、独立変数によって説明可能な変動にムラが出る



誤差分布の正規性

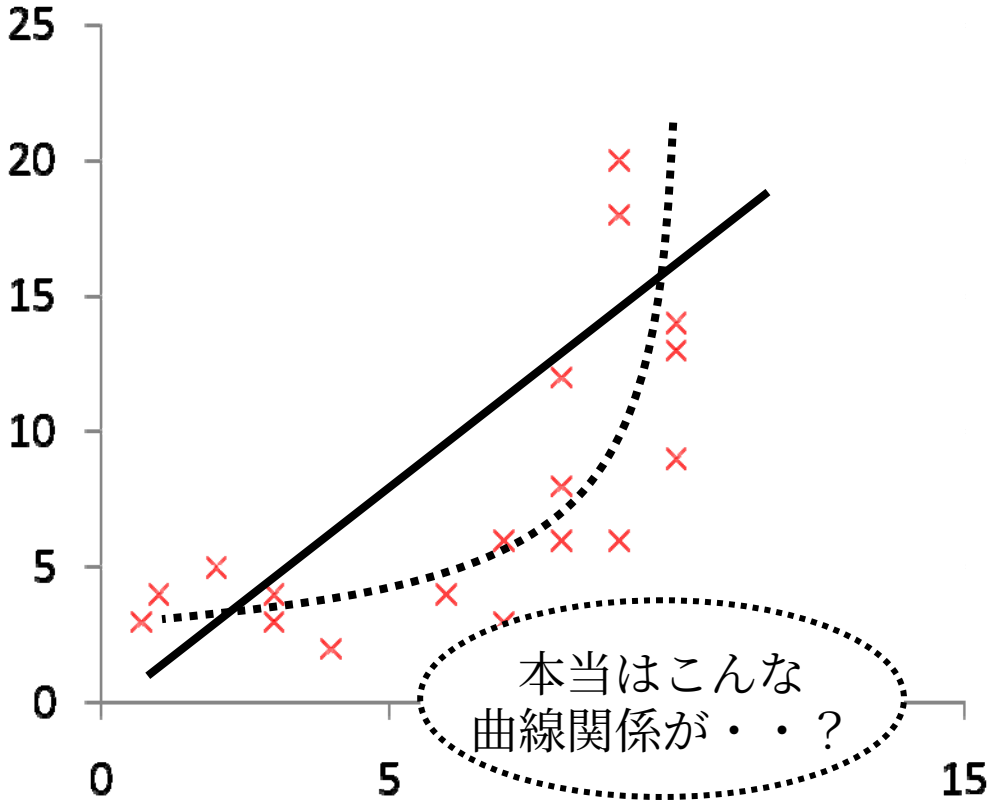
- 統計的な問題
  - 2分散の比は両変数が正規分布に従う場合にのみ*F*分布に従うと仮定できる
  - この仮定が満たされない場合、有意性の判定が不正確になる

- 散布図上の問題
  - 実際には存在しない関係を見出す可能性
    （例、切断効果）
  - 変数間の関係がモデルに投入していない変数によって異なる可能性
    （例、交互作用）



線形性（加法性）

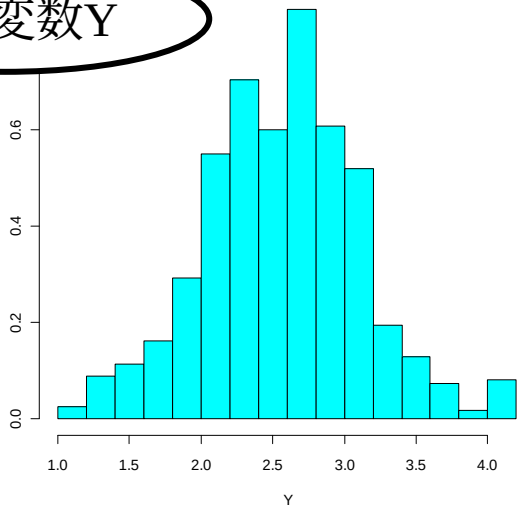
- 分析で扱ったモデル式が線形でも、実際の分布（実際の変数間の関係）は非線形の可能性もある
  - 実際のデータに対して
    適切なモデル式を用いて
    分析したとは言い難い



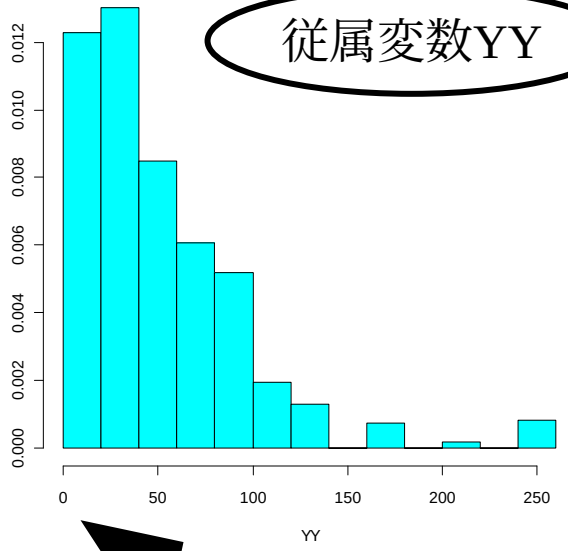
これから使うデータ

- この後の図では以下の変数およびモデルを使う

従属変数Y

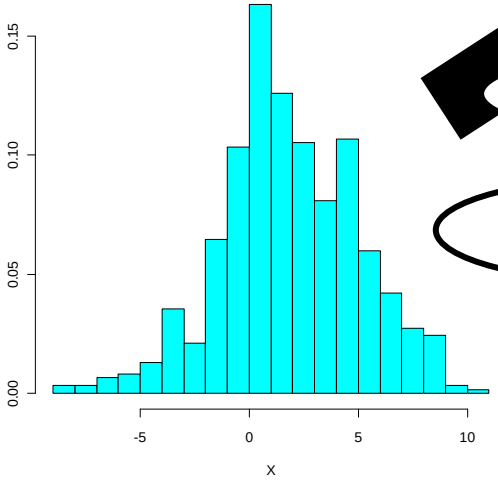


従属変数YY



モデルA (Y = αX + e_A)

モデルB (YY = βX + e_B)

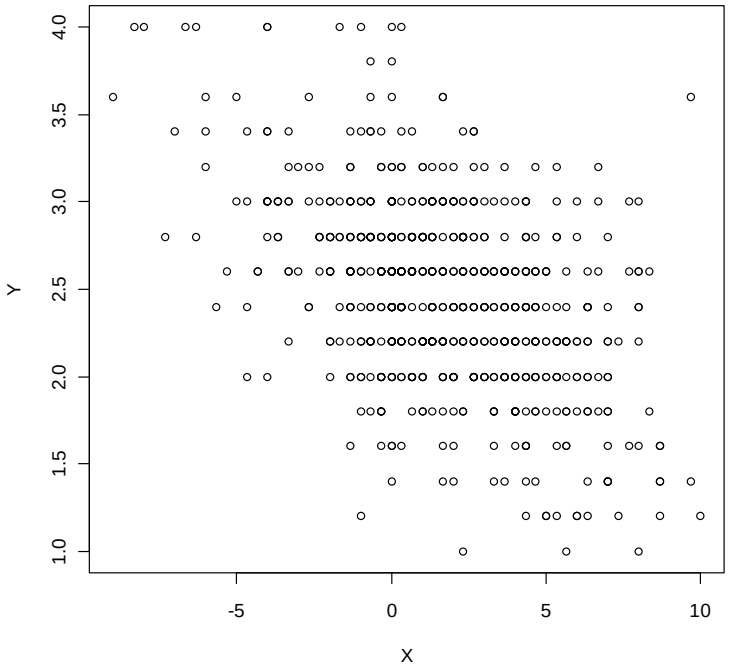


独立変数X

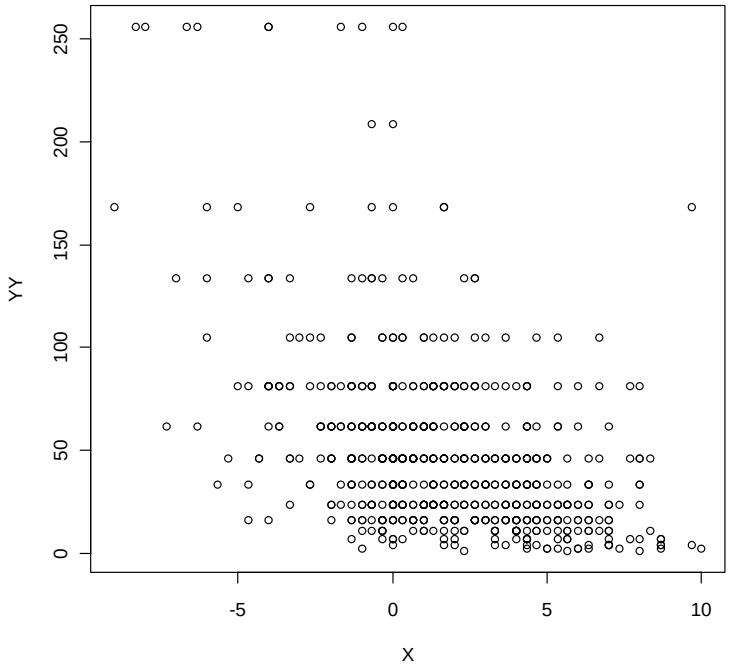
これから使うデータ

- ちなみに散布図は・・・

モデルA ($Y = \alpha X + e_A$)



モデルB ($YY = \beta X + e_B$)



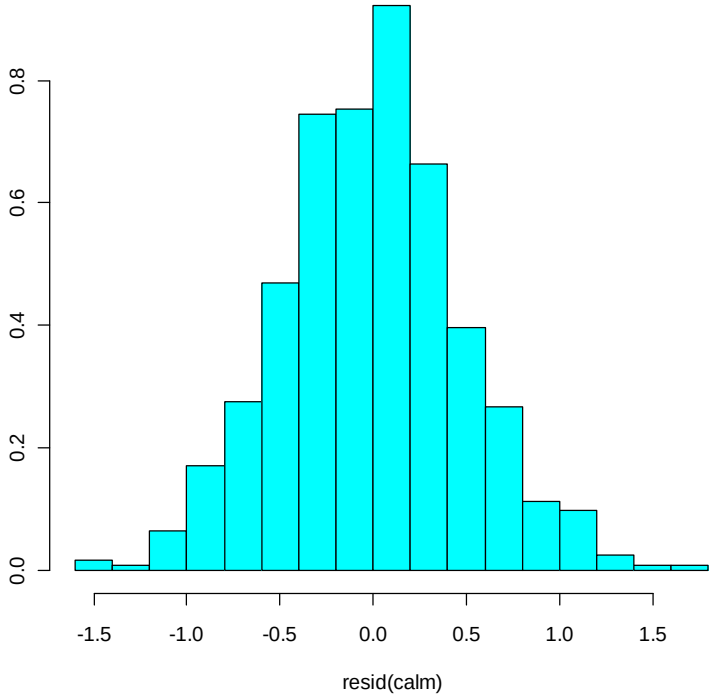
- 分析上は、いずれも有意であり、変数間には負の関係が

分布の正規性：残差のヒストグラム

- 残差
  - 適合値（モデル上の値）と、実際のデータの差
  - 定義上、残差の平均は**0**になる

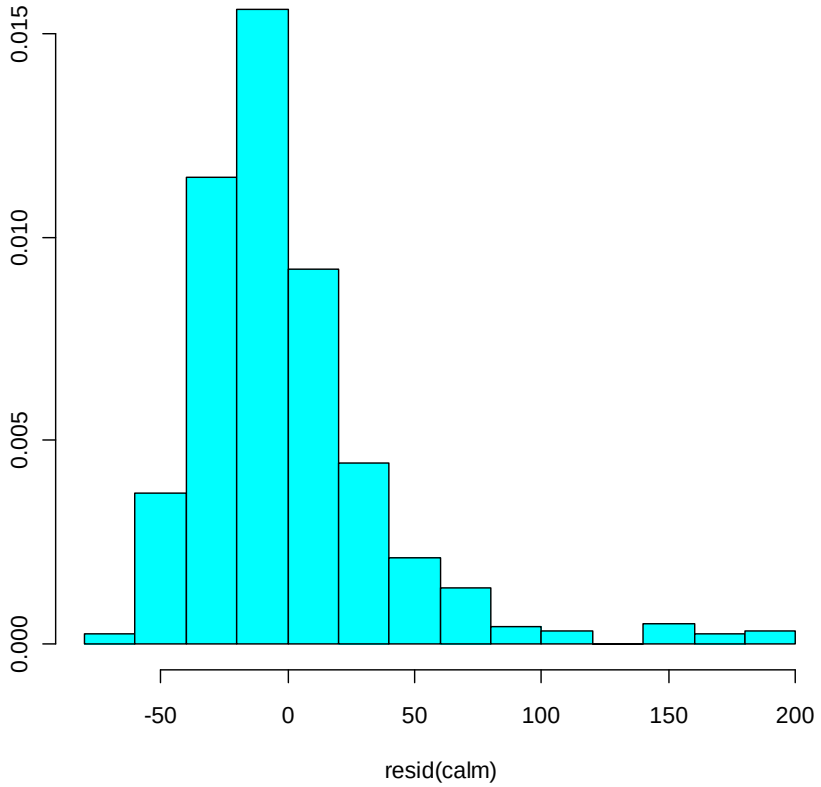
    （モデルは各データ点からの距離が最小(　0)になるように求められるため）
  - さらに残差を標準化することで、各データ点がどれだけモデルから離れているかが判断できる。

- 右図はモデルAの残差のヒストグラム
- ほとんどの値が平均値0を中心に、SD±1の間に正規的に分布しているように見える



分布の正規性：残差のヒストグラム

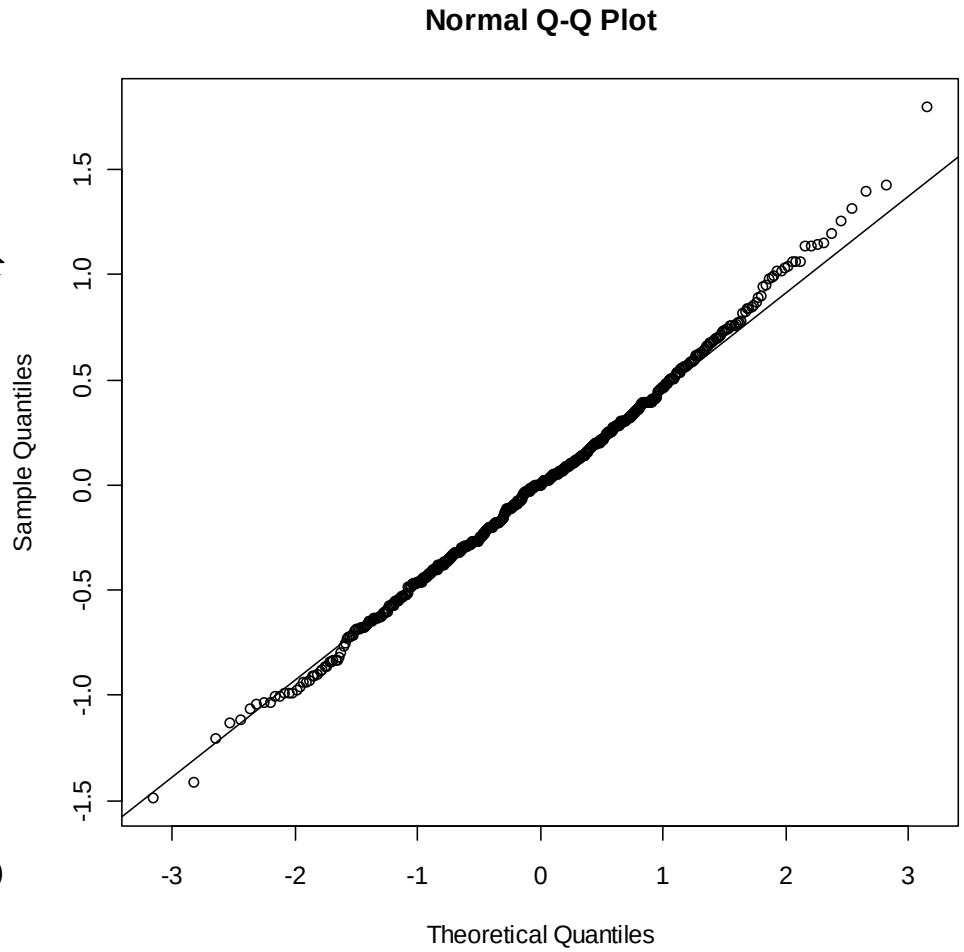
- 右図はモデルBの残差のヒストグラム
- 分布はやや右に裾を延ばしている
- 残差の分布は元の従属変数の分布に近いかたちになる



分布の正規性：正規確率プロット

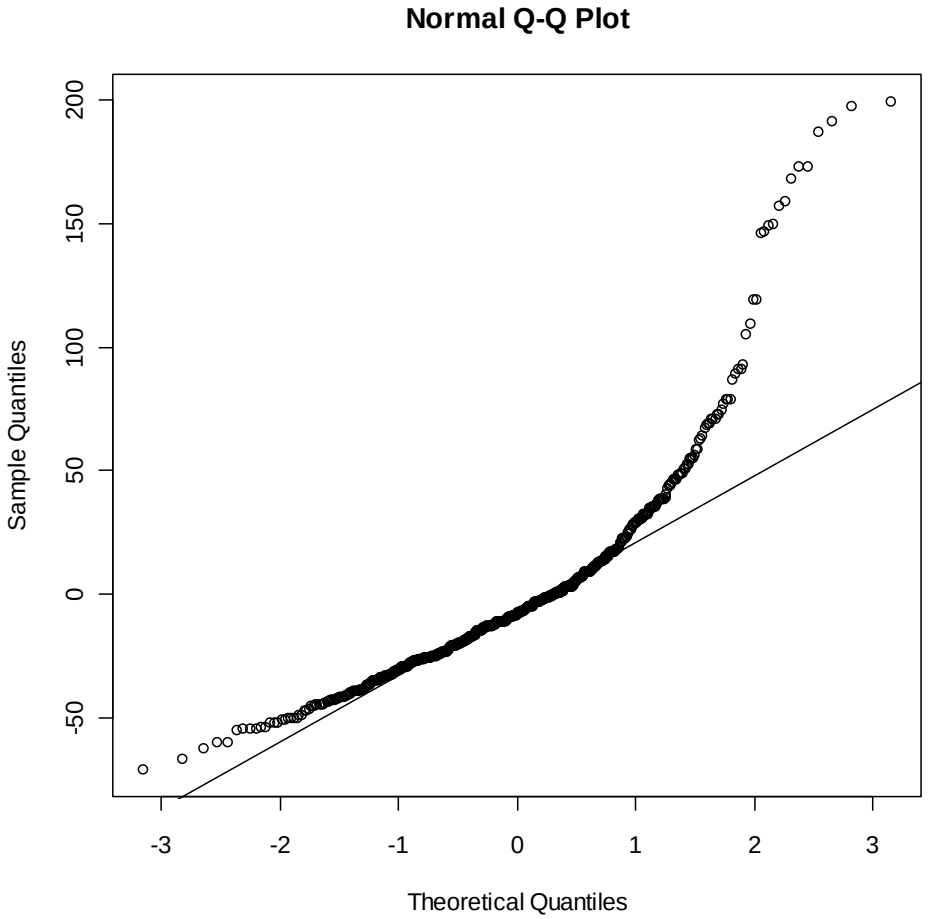
- 残差（y軸）と残差の期待値(x軸)を図示したもの
  - 標準化された残差を大きさの順に並べてみる
  - 標準化された残差が正規分布に従う場合に取り得る値の期待値を計算することができる
  - 実際の残差が正規分布に従っているならば、両値の相関が高くなるはず

（右図はモデルA）



分布の正規性：正規確率プロット

- 右図はモデルB
- 順位の高い方で、実際の残差が期待値よりも大きな値を持っているために、凹状の分布になっている

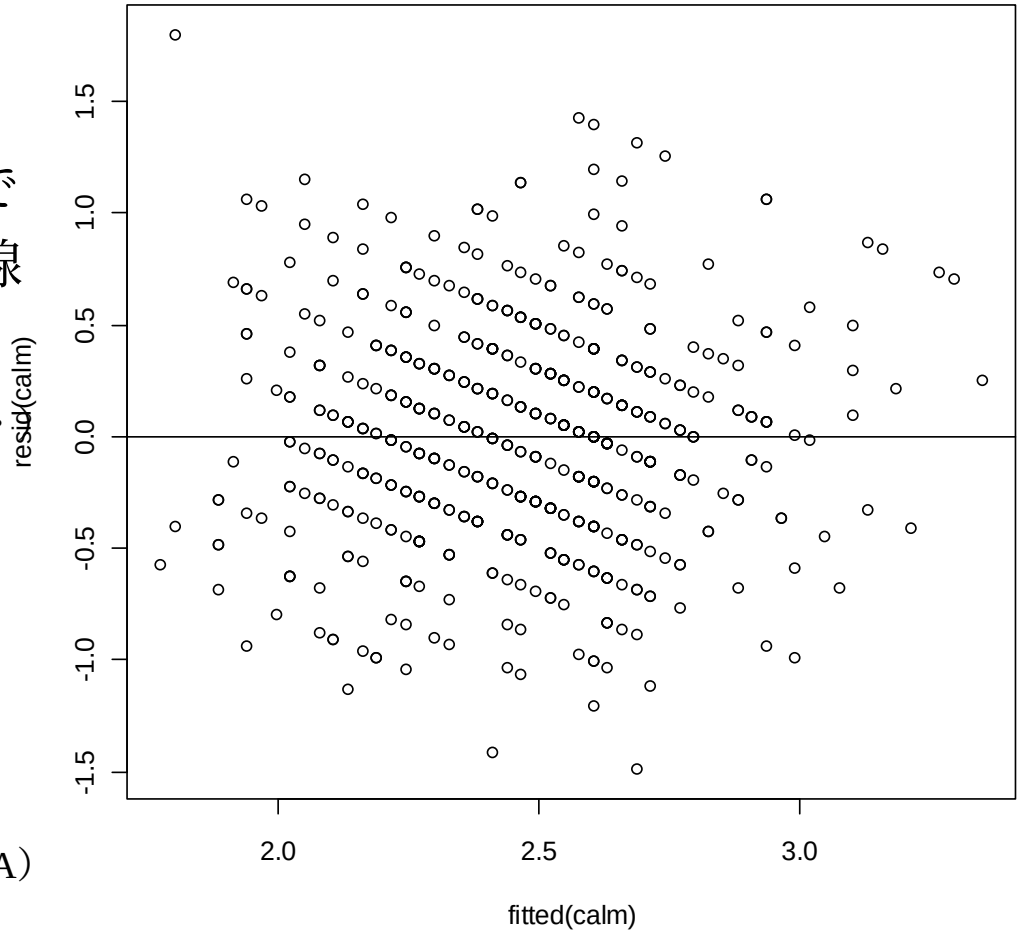


線形性と分散の均一性：適合値に対する残差プロット

- 適合値をx軸、標準化残差をy軸にとり、図示したもの
  - モデルがデータに完璧に当てはまっていたならば、誤差は全く存在しないので残差はすべてx = 0の水平線上にプロットされる

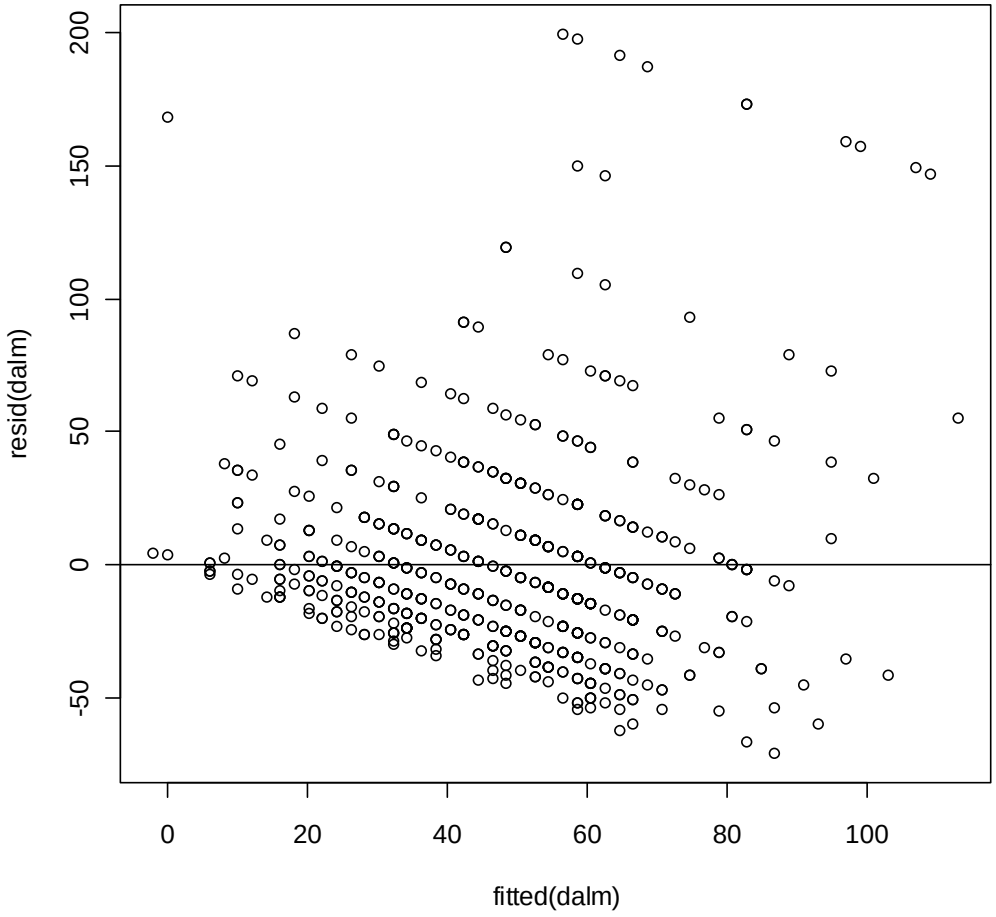
    （まずありえないが・・・）
  - 誤差変動が均一ならば、残差はx = 0の水平線の上下に均等に散らばってプロットされる

    （右図はモデルA）



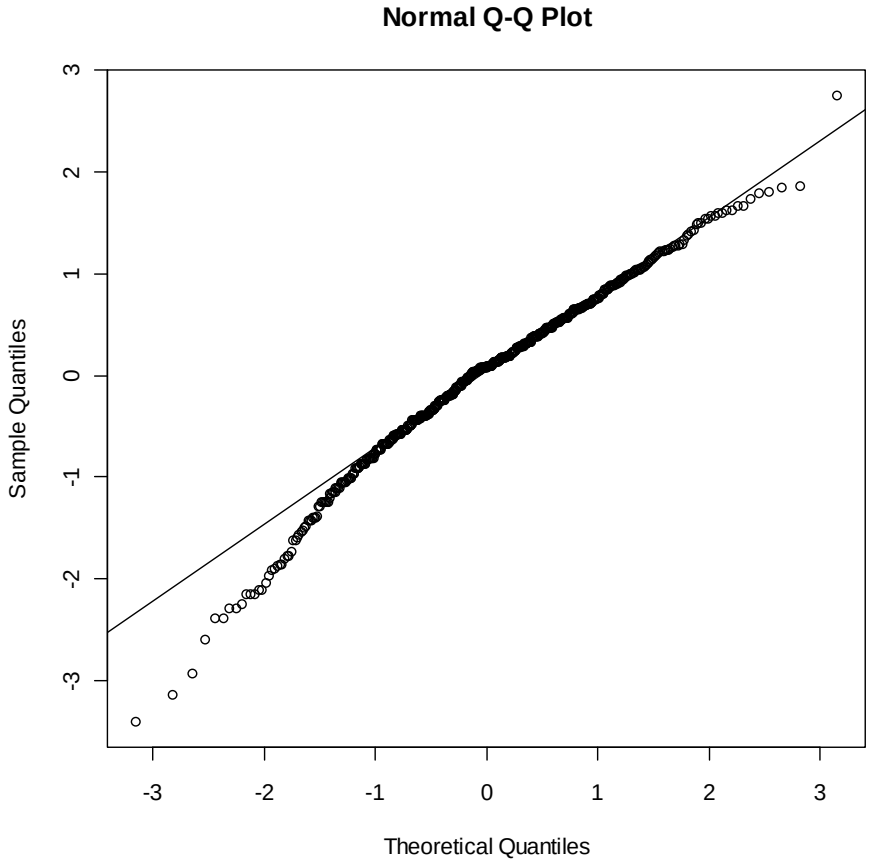
線形性と分散の均一性：適合値に対する残差プロット

- 右図はモデルB
- 適合値が大きくなるにつれて分散が増大しているように見える
- 何らかの変数が適合値と相乗的に関連している可能性



分散の正規性、均一性：変数の変換

- モデルBを、$\log(YY) = \beta X + e_B$のように従属変数を対数変換
  - 右図は変換後の正規確率プロット
  - 変換前と比べると、直線的な関係に近づいていた
  - 他にも、逆数変換・平方根・指数等の変換が考えられる
  - **ただし、その変換をして分析することの妥当性は考慮しなければならない**

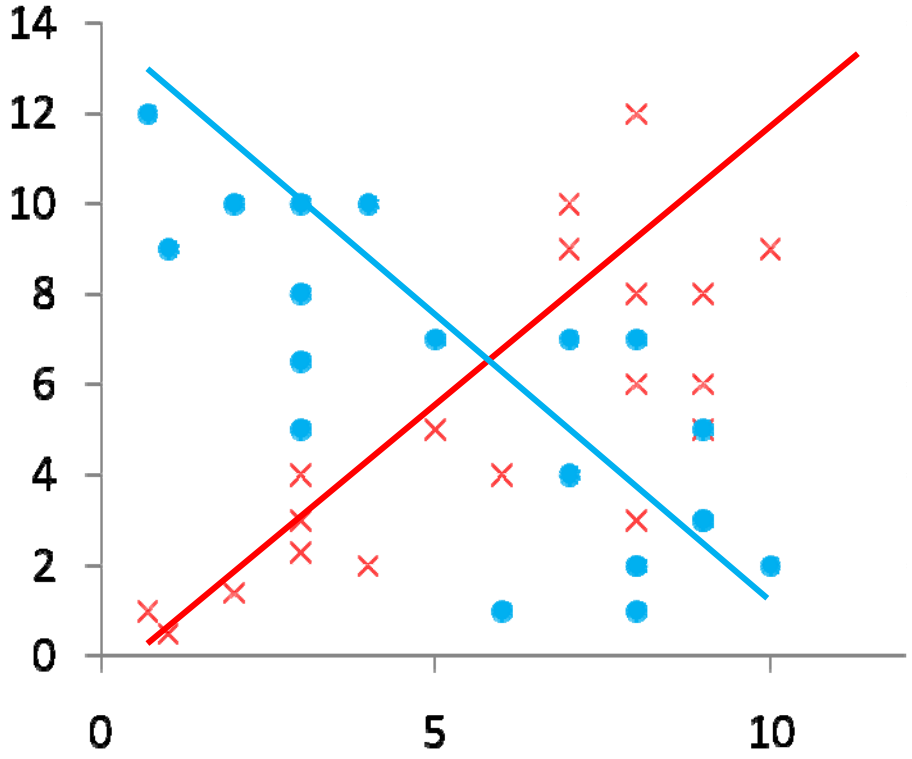


分散の均一性、線形性：交互作用

- 変数間の関係が非線形の場合、交互作用が有効なときも
- カテゴリ変数と連続変数の交互作用
  - 例）従属変数Y(連続)、独立変数A(連続)、B(カテゴリ)

モデル式　Y = A + B + A*B・

  - 「交互作用が有意」な場合、変数Aと変数Yの関係が、Bの値によって異なることを示す
  - つまり、図示すると右図のように、カテゴリ変数Bの値で傾きが異なる図になる



分散の均一性、線形性：交互作用

- 連続変数同士の交互作用
  - 例）従属変数Y(連続)、独立変数A(連続)、B(連続)

    モデル式　Y = A + B + A*B

  - 「交互作用が有意」な場合、変数Aと変数Yの関係がBの値によって異なることを示す
  - 図示するのは困難だが、平均+1SD群と平均-1SD群で高低群に分割して作図することもある（ようだ）

  - 例えば・・・
    - ある樹木の体積を、高さと直径を使って予測するとする
    - 高さ×直径の交互作用が有意ならば、直径が大きくなるほど高さ1単位分のもつ体積への影響の度合いが大きくなる（またはその逆）ことを示す

まとめ

- 重回帰分析を行った後・・・
  - データが重回帰分析を行うために必要な仮定を満たしているかチェックしてみる
  - 正規確率プロット・適合値に対する残差プロットなど
  - 他にもいろいろあるみたいです（例、Cook の距離など）

- 満たしていなかった場合
  - 別のモデルを考える（交互作用も含め）
  - 変数に変換を加えてみる（対数・平方根・指数など）
  - ただし、最も適合するモデルを得ることも大事だが、解釈可能であることも必要

実際にやってみる

① 書籍中の練習問題用に作られたデータから重回帰分析を行い、今日見てきた診断方法を試してみる

② （あとででも）自分のデータについても試してみる

- 以下は①について説明する

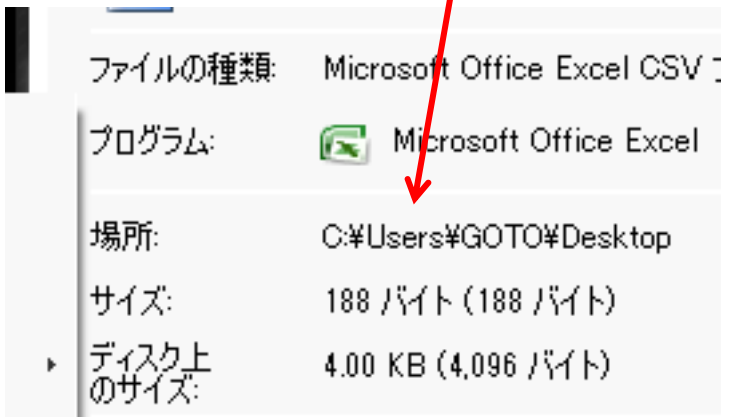
①練習問題（『一般線形モデルの～』より改変）

- ある研究で策定された健康維持計画において、19人の学生の体重が測定され、また、体脂肪率の推定値が求められた

- “goto.csv”
  - FAT　　体脂肪率（%）
  - WEIGHT　　体重（kg）
  - SEX　　性別（1 = 女性, 2 = 男性）

- 問題①
  - 体重は体脂肪率を
  - 予測するだろうか？

“ファイル上で右クリック→プロパティ”で、ファイルの場所を確認



Rを使った回帰分析（一般線形モデル）

- まずはデータを読み込む

```
> FATdata <- read.csv(“ファイルの場所/goto.csv”)
```

- 読みこめているかどうかを確認・・・

```
> FATdata
```

- いちいちデータフレーム名を書くのはめんどくさいので省略

```
> attach(FATdata)
```

Rを使った回帰分析（一般線形モデル）

- とりあえず単回帰分析してみる

```
> fat.lm <- lm(FAT ~ WEIGHT)
> # lm(Y ~ X)は、"Xを独立変数、Yを従属変数とした一般線形モデル
> # を作る"という命令
```

  - 体重を独立変数、体脂肪率を従属変数とした単回帰モデルを"fat.lm"と定義

```
> summary(fat.lm)
> # summary()で分析結果の要約を返す
> # ちなみにanova()で()のモデルについて分散分析をします
```

  - 分析結果の要約を返す

Rを使った回帰分析（一般線形モデル）

- 結果要約

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.2715 -2.7508  0.1906  1.9699  6.6871

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.88558    4.67036   5.757 2.33e-05 ***
WEIGHT       0.02069    0.06414   0.323    0.751
---

Residual standard error: 3.574 on 17 degrees of freedom
Multiple R-squared: 0.006081,  Adjusted R-squared: -
0.05238
F-statistic: 0.104 on 1 and 17 DF,  p-value: 0.751
```

β

R^2

Rを使った回帰分析（一般線形モデル）

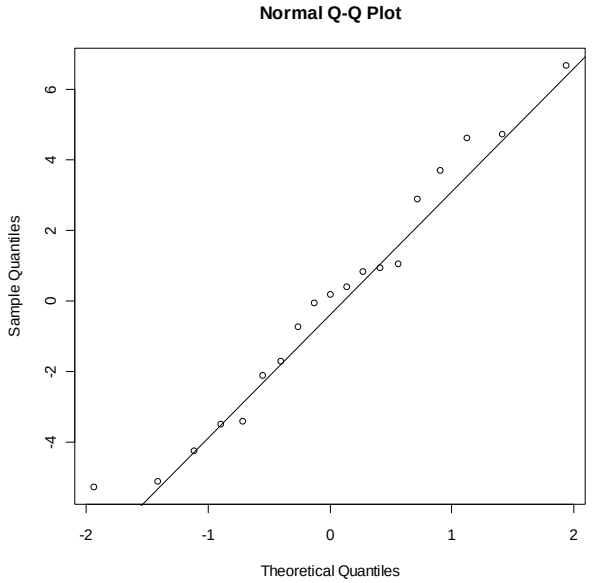
- 分析の結果、$\beta = .02, p = .75$となり、体重は体脂肪を予測するとはいえなかった
- ここで、回帰分析の仮定が満たされているかを検査してみる

正規確率プロット

- 正規確率プロットを見てみる

```
> qqnorm(resid(fat.lm))
> qqline(resid(fat.lm))
> # resid()は()内のモデルの残差を取り出す
> # qqnorm()で正規確率プロット、qqline()でプロット上に直線を引く
```

- 残差の分布はそれほど正規分布から離れてはいないように見える

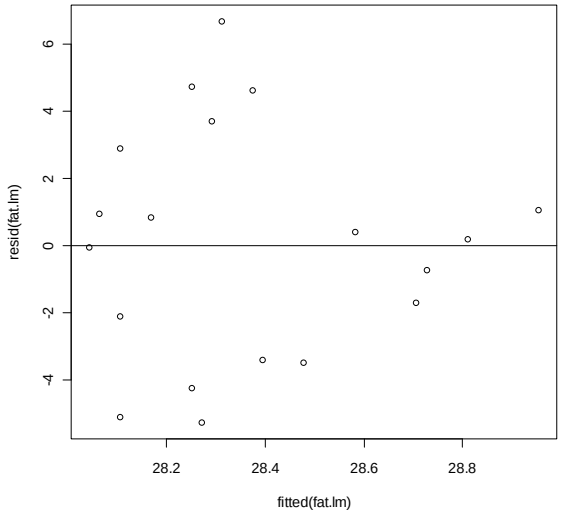


適合値に対する残差のプロット

- 適合値に対する残差のプロットを見てみる

```
> plot(fitted(fat.lm), resid(fat.lm))
> abline(h = 0)
> # fitted()は()内のモデルの適合値を取り出す
> # plot(x, y)で散布図を描く
> # abline(h = 0)は"y = 0の直線をひけ！"という命令
```

- 分散は均一ではなさそう
- 2つの集合があるようにも
  見える・・・？
  （実は普通に散布図を取っても
  こんなかたちになっている）



①練習問題

- 問題②
  - 体重が体脂肪率を予測する度合いは性別によって異なっているか？

Rを使った回帰分析（一般線形モデル）

- 交互作用を入れて重回帰分析してみる

```
> fatx.lm <- lm(FAT ~ WEIGHT + SEX + WEIGHT*SEX)
> # lm(Y ~ X1 + X2 + …)で、独立変数を複数投入できる
> # 今回、交互作用は独立変数同士の積で投入するのでX1*X2と表す
> # 書き表し方は複数ある(例、lm(FAT ~ (WEIGHT + SEX)^2))
```

  - 体重、性別、体重と性別の交互作用を独立変数、体脂肪率を従属変数とした単回帰モデルを"fatx.lm"と定義

```
> summary(fatx.lm)
> anova(fatx.lm)
```

  - 分析結果の要約と分散分析の結果を返す

Rを使った回帰分析（一般線形モデル）

- 分散分析表（結果要約は省略）

```
Response: FAT
            Df  Sum Sq Mean Sq F value    Pr(>F)
WEIGHT       1   1.328   1.328  0.6611   0.42889
SEX          1 176.098 176.098 87.6467 1.181e-07 ***
WEIGHT:SEX   1  10.857  10.857  5.4039   0.03454 *
Residuals   15  30.138   2.009
---
```

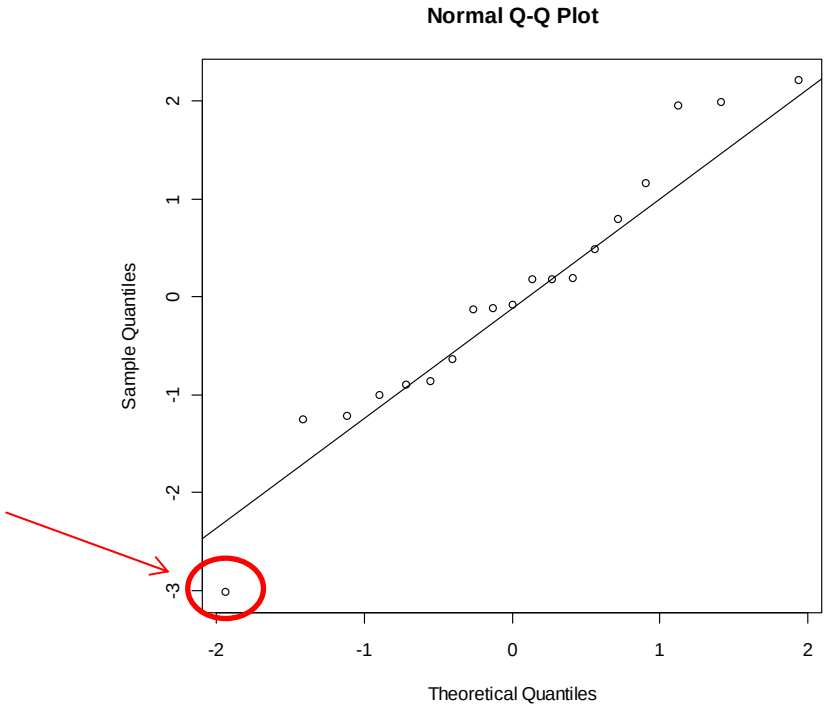
- 交互作用が有意 (F (1, 15) = 5.40, p < .05)
  - モデル検査は・・・？

正規確率プロット

- 正規確率プロットを見てみる

```
> qqnorm(resid(fatx.lm))
> qqline(resid(fatx.lm))
```

- 残差の分布はそれほど正規分布から離れてはいないように見える
- この値が外れ値気味な印象も受けるが・・・

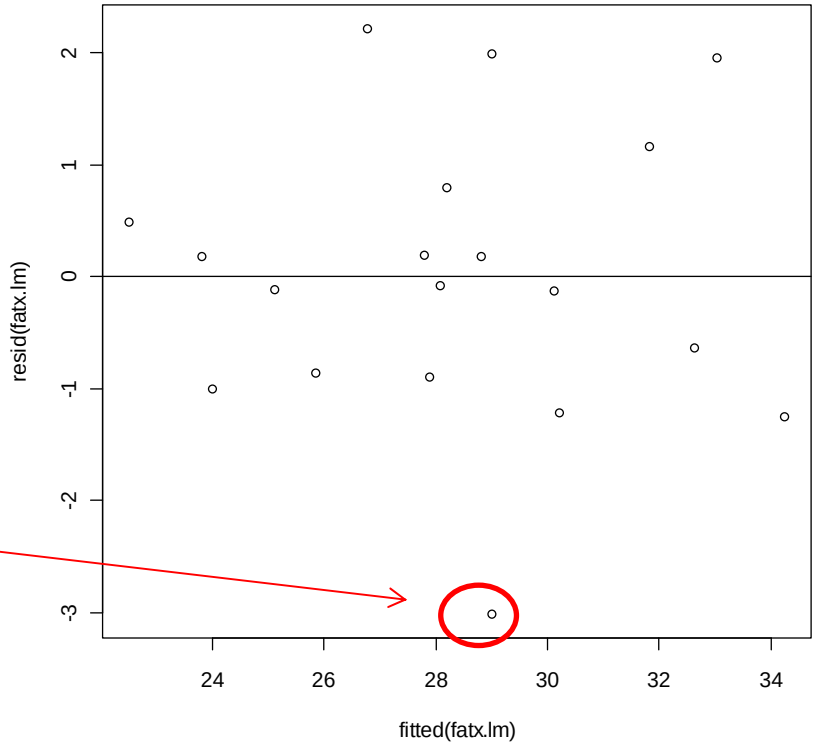


適合値に対する残差のプロット

- 適合値に対する残差のプロットを見てみる

```
> plot(fitted(fatx.lm), resid(fatx.lm))
> abline(h = 0)
```

- 分散は均一性は前のモデルに比べて保たれてるように見える
- やはりこの値は外れ値気味な印象も・・・



Rを使った回帰分析（一般線形モデル）

- 性別ごとの分析へ

```
> female <- subset(FATdata, SEX == "1")
> male <- subset(FATdata, SEX == "2“)
> # subset(DF, 条件)で、条件に合うデータのみであらたなDFを作成
> # =ではなく==を使い、数値は””で囲む
```

  - 男性のみデータフレームと女性のみデータフレームを作成

```
> attach(female) # 男性はattach(male)
> fatf.lm <- lm(FAT ~ WEIGHT)
> coef(fatf.lm)
> # coef()はモデルの回帰係数を求める
```

  - 性別ごとに回帰係数を求める

Rを使った回帰分析（一般線形モデル）

- 結果・・・

```
> coef(fatf.lm)
(Intercept)      WEIGHT
  5.2396694   0.4028926

> coef(fatm.lm)
(Intercept)      WEIGHT
 11.5709579   0.1855043
```

- 女性：体脂肪率 = 0.40 * 体重 + 5.24
- 男性：体脂肪率 = 0.19 * 体重 + 11.57
  - 男女で体重と体脂肪率との関係は異なっていた

Rを使った回帰分析（一般線形モデル）

- ちなみに

```
> layout(matrix(1:4, 2, 2, byrow=TRUE))
> plot(fatx.lm)
> # plot(モデル)でモデル検査の図がいろいろ出てきてくれます
```

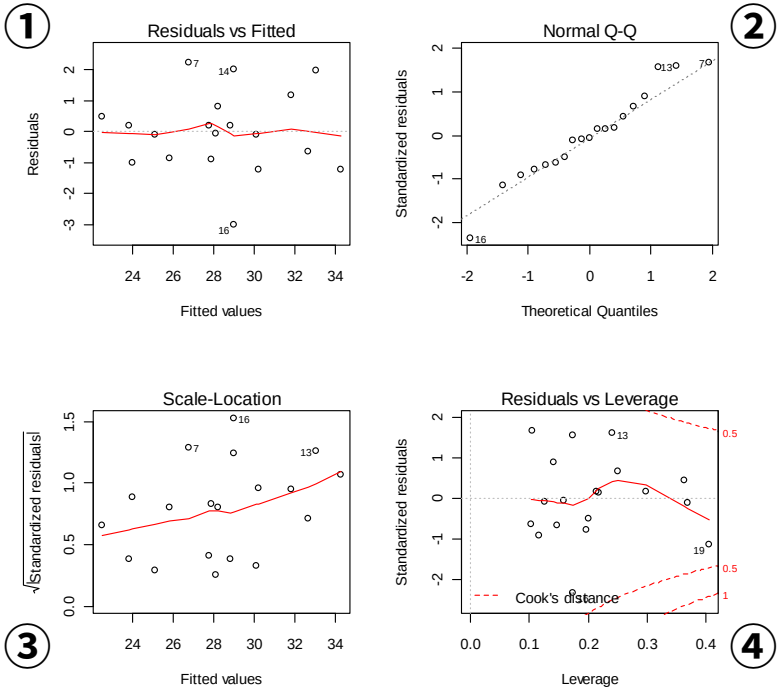
- 左上から横に、

① 適合値に対する残差プロット

② 正規確率プロット

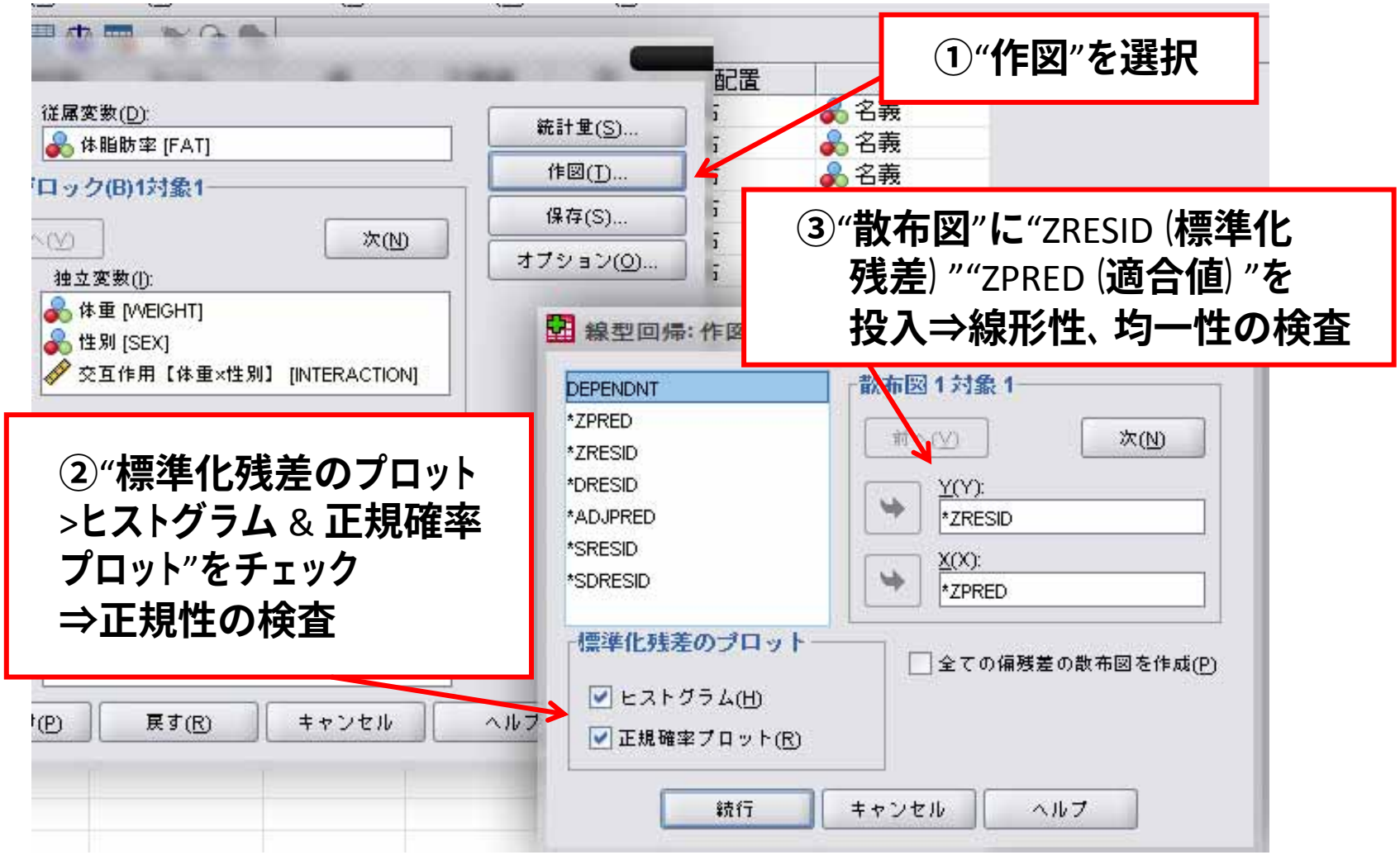
③ 適合値に対する標準化された残差プロット

④ Cook の距離



おまけ

- モデル検査：SPSSの場合



おまけのおまけ

- モデル検査：SPSSの場合：シンタックスの場合

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT FAT
  /METHOD=ENTER WEIGHT
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HIST(ZRESID) NORM(ZRESID).
```

散布図

正規確率プロット

- 逐一シンタックスを保存しておくと、分析の手順を後から振り返ることができるので便利です

おまけのおまけ

- 交互作用：SPSSの場合：シンタックスの場合

交互作用として積の変数作成

```
compute INTERACTION = WEIGHT * SEX.
variable labels INTERACTION '交互作用【体重×性別】'.

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT FAT
  /METHOD=ENTER WEIGHT SEX INTERACTION
```

交互作用投入！

参考資料

- グラフェン, A. & ヘイルズ, R. (2007). 一般線形モデルによる生物化学のための現代統計学　―あなたの実験をどのように解析するか―. 野間口謙太郎, 野間口眞太郎 (訳). 共立出版. [Grafen, A. & Hails, R. (2002). *Modern Statistics For The Life Sciences*. Oxford University Press.]
- データ: (http://www.oup.com/uk/orc/bin/9780199252312/)

- 青木繁伸. (2009). Rによる統計解析. オーム社.

- R-Tips (http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html)