環境統計学ぷらす

第3回
一般線形モデル・交互作用

高木　俊
shun.takagi@sci.toho-u.ac.jp

2013/11/07

予定

- 第1回：　Rの基礎と仮説検定
- 第2回：　分散分析と回帰
- 第3回：　一般線形モデル・交互作用
- 第4回：　一般化線形モデル・モデル選択
- 第5回：　一般化線形混合モデル
- 第6回：　多変量解析

今日やること

- R操作編
  - 連続変数とカテゴリカル変数
  - パッケージの読み込み
- 統計編
  - 二元配置分散分析
  - 重回帰分析
  - 共分散分析
- 表現編

Rの操作

- 連続変数とカテゴリカル変数
- パッケージのインストール

R上でのデータ型

- read.csv()などでデータフレームとして読み込んだ場合、数字は連続変数、文字列はカテゴリカル変数として読み込まれる

```
> data3.1<- read.csv("data3.1.csv",T)
> is.numeric(data3.1$id)                    #型がnumeric（実数）か？
[1] TRUE
> is.factor(data3.1$id)                     #型がfactor（因子）か？
[1] FALSE
> is.numeric(data3.1$nutrients)
[1] FALSE
> is.factor(data3.1$nutrients)
[1] TRUE
```

基本的に、因子（カテゴリカル変数）として読ませたい場合は文字列でデータを作っておくのが無難

型変換

- 数字をfactor(順序なし因子)として扱いたい場合は、型の変換を行う

```
> data3.1$id<- as.factor(data3.1$id)      #as.factorで因子に変換
> is.numeric(data3.1$id)
[1] FALSE
> is.factor(data3.1$id)
[1] TRUE
```

numeric→factorで型変換が行われることで、
plot()　散布図 → 箱ひげ図
lm()　　回帰　→ 分散分析
など実行内容が変化するので注意

パッケージのインストール

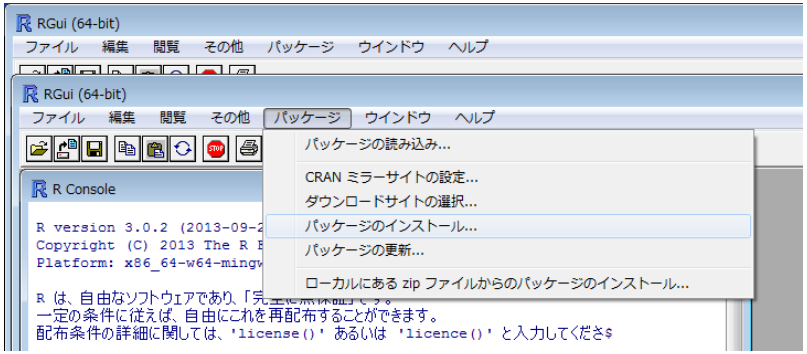
- Rでは最初に用意されている基本パッケージに加え、有志によって作られた様々なパッケージを利用することができる
- 高度な解析を行う場合には、目的に応じたパッケージをインストールする必要がある

（例）

パッケージ名	内容
car	回帰・分散分析系の拡張機能
MASS	一般化線形モデル（負の二項分布など）
lme4, glmmML	一般化線形混合モデル（二項分布・ポアソン分布）
glmmADMB	一般化線形混合モデル（負の二項分布・ZIP・ZINB）
MuMIn	モデルアベレージング・モデル選択
vegan	多変量解析・生物多様性解析

インストール方法

- Rのメニューバーから
  「パッケージ」＞「パッケージのインストール」



- ミラーサイトを選択。Japan(Tokyo)でOK（特に変わりない）
- パッケージ一覧から今回は「car」を選択

一度インストールしたパッケージはlibrary(パッケージ名)
もしくは「パッケージの読み込み」から呼び出せる

注意点・その他

- パッケージの中の関数一覧はlibrary(help="パッケージ名")から参照。ダメな場合help(パッケージ名)から飛べる？
- Rのバージョンによって、インストールできるパッケージやそのバージョンも異なる。使いたいパッケージが一覧にない場合や、使ってみたけど出力表示が他の人と異なる場合、Rの最新版(場合によっては古い版)をインストールし直すことで解決できることも
- デフォルトのパッケージ以外を使った場合は、パッケージ名も方法に書いたほうが良いかも

(例)
R3.0.2 (R Core Team 2013)およびveganパッケージ(Oksanen 2013)を用いた

一般線形モデル

一般線形モデル

- 回帰も分散分析も

$$y_{ij} = \beta_0 + \sum \beta_i \times x_{ij} + \varepsilon_{ij}$$ の形で記述できる

Rでは　lm(Y~X)

- 説明変数を2個以上に増やした場合も同様の式で表すことができる

（例）

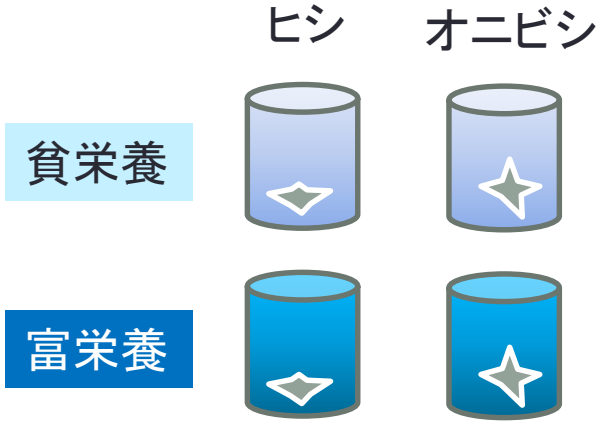
植物現存量 $= \beta_0 + \beta_1 \times$ 光の強さ $+ \beta_2 \times$ 土壌水分量 $+$ 誤差
溶存酸素 $= \beta_0 + \beta_1 \times$ ヒシの有無 $+ \beta_2 \times$ 上層or下層 $+$ 誤差

Rでは　lm(Y~X1+X2+‥‥)

2元配置分散分析

- 説明変数（カテゴリカル型）が2種類ある場合の分散分析

data3.1　ヒシ・オニビシ培養実験データ



×4繰り返し

最終バイオマスに対して
- 種の違い
- 栄養塩濃度の違い

は影響するか？

- 要因Aと要因Bのそれぞれの効果を見たい場合
- 要因Aを考慮した上で要因Bの効果を見たい場合
- 要因Aと要因Bの相互作用を見たい場合

灌漑の種類と品種の違いは収量に影響するか

雌雄差を考慮した上で薬の処理は代謝に影響するか

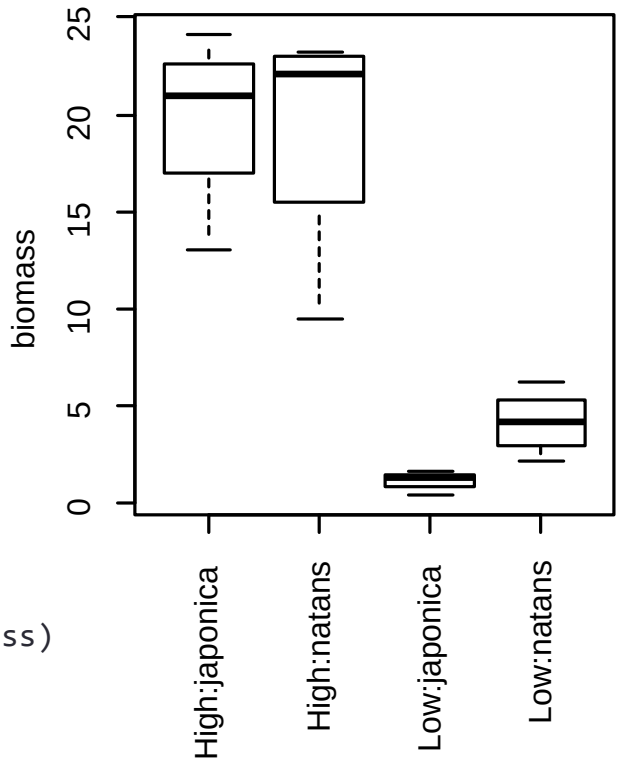
種類によって栄養に対する反応が異なるか

とりあえず図示

- 栄養塩処理（nutrients）と種類（species）の組み合わせごとにデータを示したい

```
# 要因同士をコロンでつないで新たなデータ列を定義
> data3.1$n.s<- data3.1$nutrients:data3.1$species
> data3.1$n.s
 [1] High:japonica High:japonica High:japonica
High:japonica High:natans
（中略）
Levels: High:japonica High:natans Low:japonica
Low:natans
> plot(biomass~n.s,data3.1,xlab="",las=3)
```

X軸名なし　文字の方向指定



平均値だけでよいならinteraction.plot()で手っ取り早く見られる

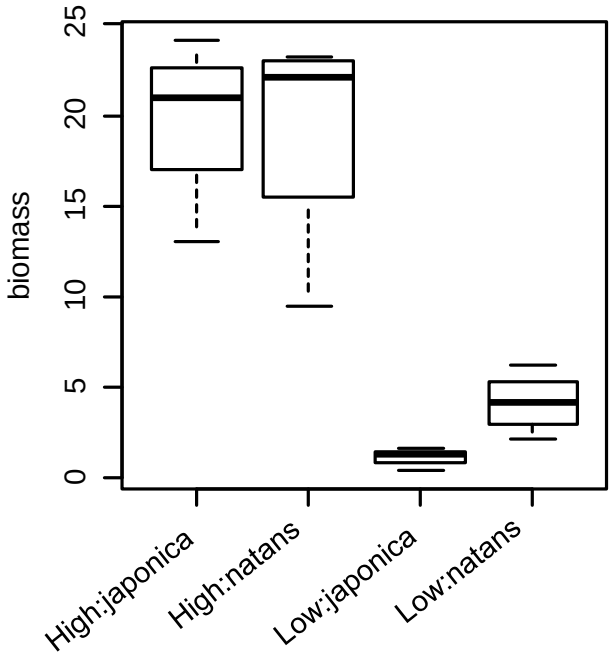
```
interaction.plot(data3.1$nutrients,data3.1$species,data3.1$biomass)
```

X軸の要因　トレースする要因　Y軸

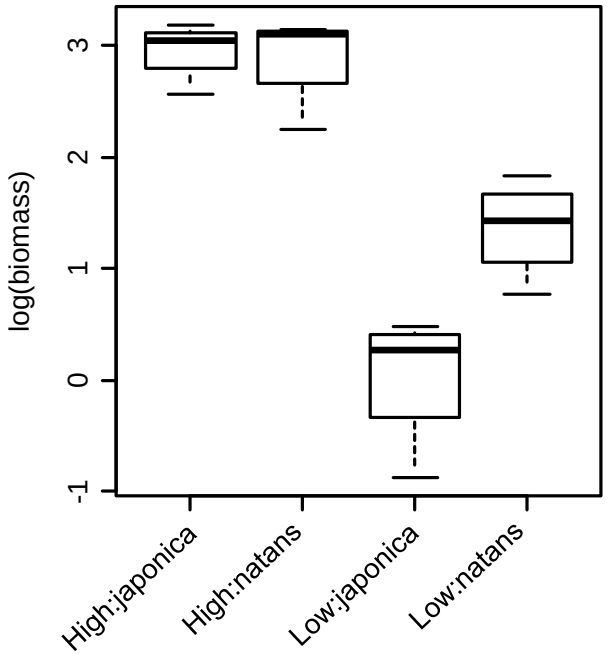
変数の変換

- データを見ると、値が小さいほどばらつき少なそう？・・・正規分布に当てはめられない
- **対数変換**したほうが正規分布への当てはまりが若干良さそう

```
plot(log(biomass)~n.s,data3.1,xlab="",las=3)
```



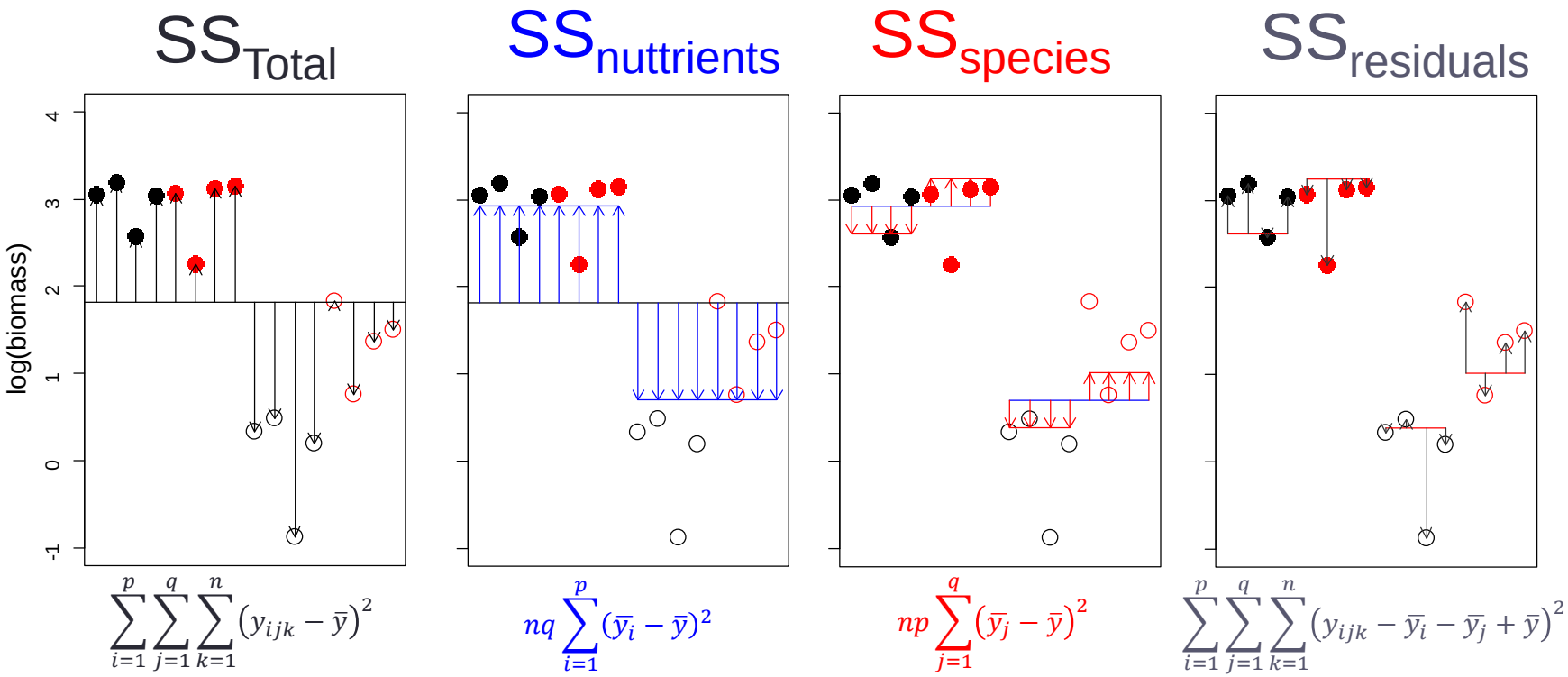
対数変換



モデルの記述と平方和の分割

- 一元配置の時と同様に線形モデルで記述
- 各要因で説明できるばらつき（平方和）に分割

$$\log(\text{バイオマス})_{ijk} = \text{栄養塩の効果}_i + \text{種の効果}_j + \text{誤差}_{ijk}$$



分散分析

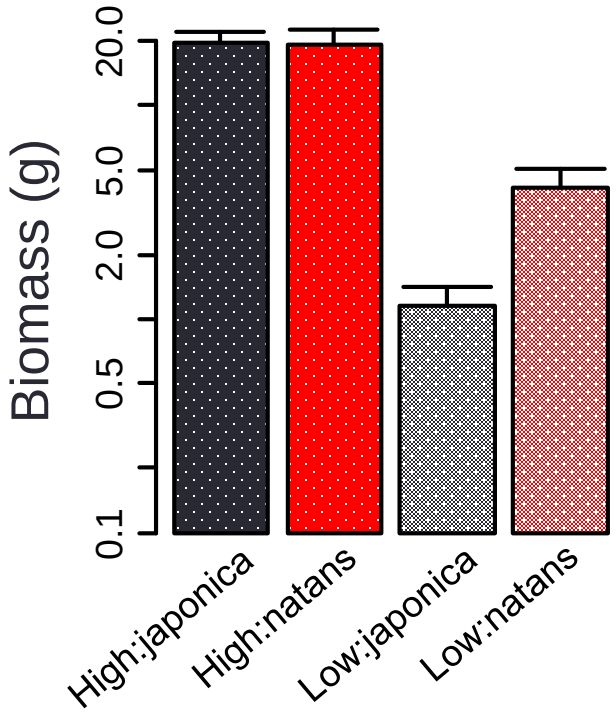
```
model3.1<- lm(log(biomass)~nutrients+species,data3.1)
anova(model3.1)
```

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
要因	自由度	平方和	平均平方	F比	P値
nutrients	1	19.8837	$MS_{nut}=SS_{nut}/Df_{nut}$ 19.8837	MS_{nut}/MS_{res} 58.1679	<0.001
species	1	1.5868	$MS_{sp}=SS_{sp}/Df_{sp}$ 1.5868	MS_{sp}/MS_{res} 4.6419	0.051
Residuals	13	4.4438	$MS_{res}=SS_{res}/Df_{res}$ 0.3418		

栄養塩による違いは明瞭だが、種による違いはなさそう？

交互作用

- 先ほどの解析では栄養塩の効果と種の効果はそれぞれ相加的に効くと仮定
- 栄養塩によって種の効果が異なる可能性もあるのでは？
  （＝種によって栄養塩の効果が異なる）



富栄養（High）の状態では
種差がないが、
貧栄養（Low）の状態では
ヒシ（japonica）がオニビシ（natans）よりも成長が悪い？

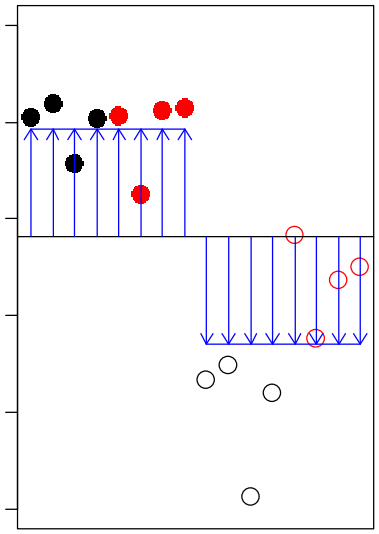
こうした非相加的関係は
交互作用（interaction）
として扱う

交互作用を考慮したモデル

先ほどのモデルではまとめて“誤差”として扱われていた

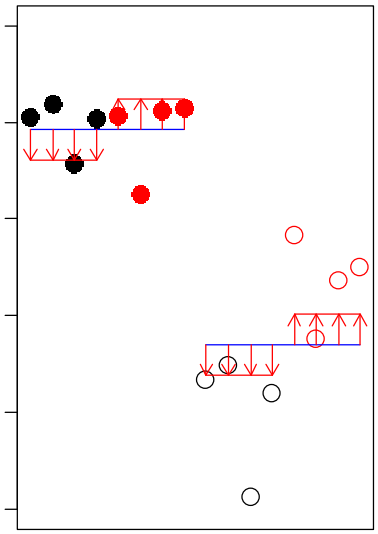
$$\log(\text{バイオマス})_{ijk} = \text{栄養塩の効果}_i + \text{種の効果}_j + \text{交互作用}_{ij} + \text{誤差}_{ijk}$$

$SS_{nuttrients}$



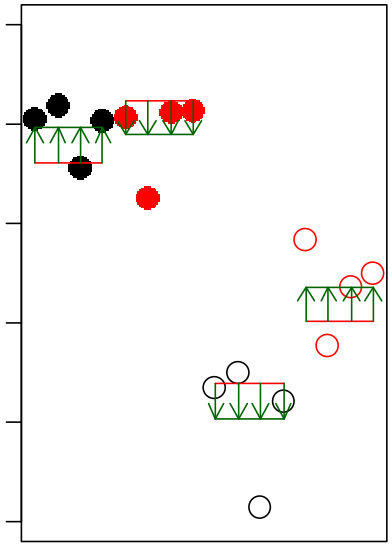
$$nq\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2$$

$SS_{species}$



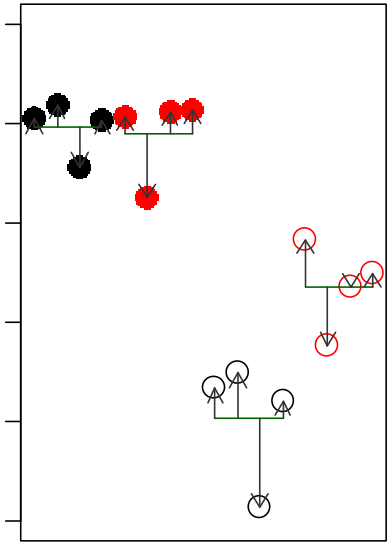
$$np\sum_{j=1}^{q}(\bar{y}_j - \bar{y})^2$$

$SS_{interaction}$



$$n\sum_{i=1}^{p}\sum_{j=1}^{q}(\overline{y_{ij}} - \bar{y}_i - \bar{y}_j + \bar{y})^2$$

$SS_{residuals}$



$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \overline{y_{ij}})^2$$

分散分析(交互作用あり)

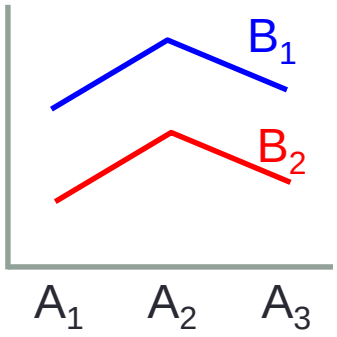
```
#交互作用込の場合はRではアスタリスクで表現
model3.1i<- lm(log(biomass)~nutrients*species,data3.1)
anova(model3.1i)
```

要因	Df 自由度	Sum Sq 平方和	Mean Sq 平均平方	F value F比	Pr (>F) P値
nutrients	1	19.8837	$MS_{nut}=SS_{nut}/Df_{nut}$ 19.8837	MS_{nut}/MS_{res} 94.8582	<0.001
species	1	1.5868	$MS_{sp}=SS_{sp}/Df_{sp}$ 1.5868	MS_{sp}/MS_{res} 7.5699	0.01756
nut:sp	1	1.9285	$MS_{int}=SS_{int}/Df_{int}$ 1.9285	MS_{int}/MS_{res} 9.2000	0.01041
Residuals	12	2.5154	$MS_{res}=SS_{res}/Df_{res}$ 0.2096		

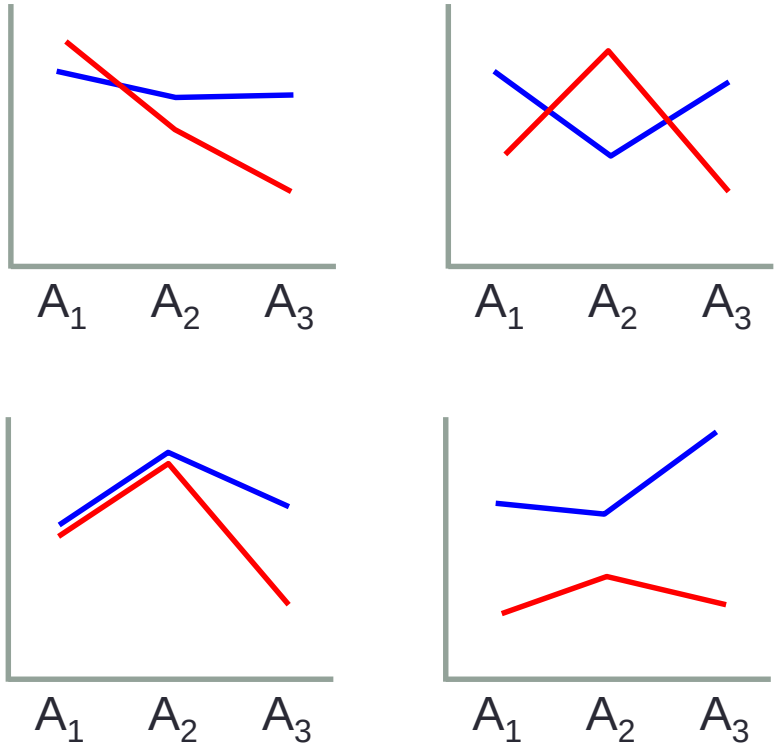
種間でのバイオマスの差は栄養塩の状態によって異なる
(＝栄養塩に対する反応は種によって異なる)

交互作用の解釈

- どういった関係にあるかは図示して確認



交互作用なし



交互作用あり

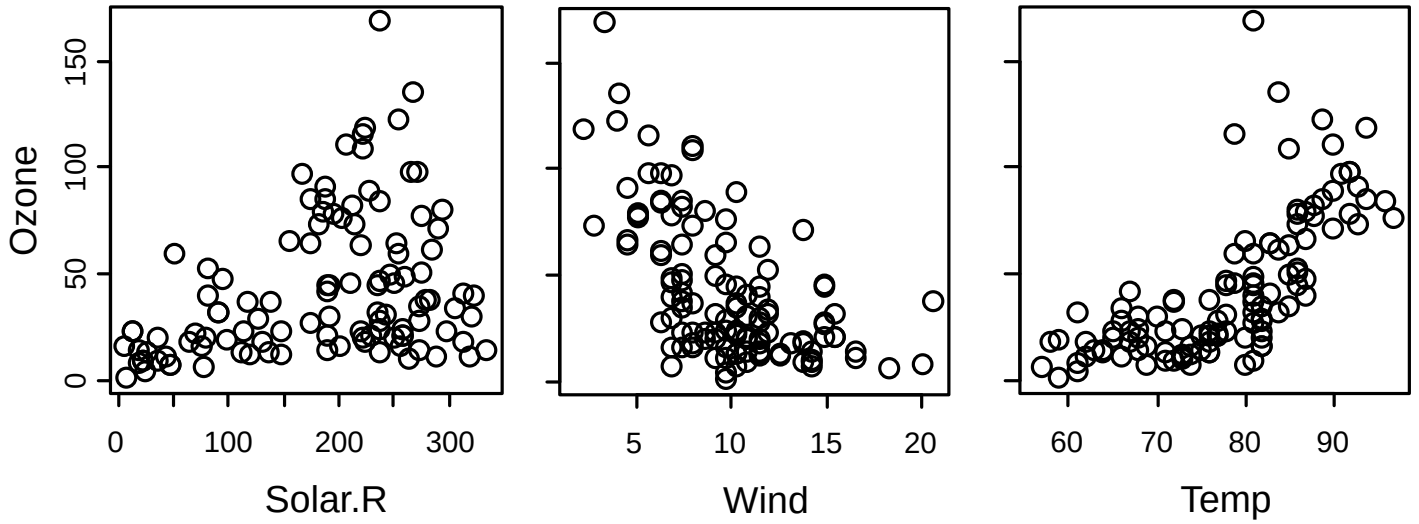
重回帰分析

- 説明変数（連続変数型）が2種以上ある場合の回帰分析

植物現存量を光条件（開空度）・土壌含水率・シカ密度で説明できるか
試験成績を前回の成績・今回の勉強量・前日の睡眠時間で説明できるか

airquality　NYの大気データ（Rの同梱データ）

?airquality　で詳細確認可能

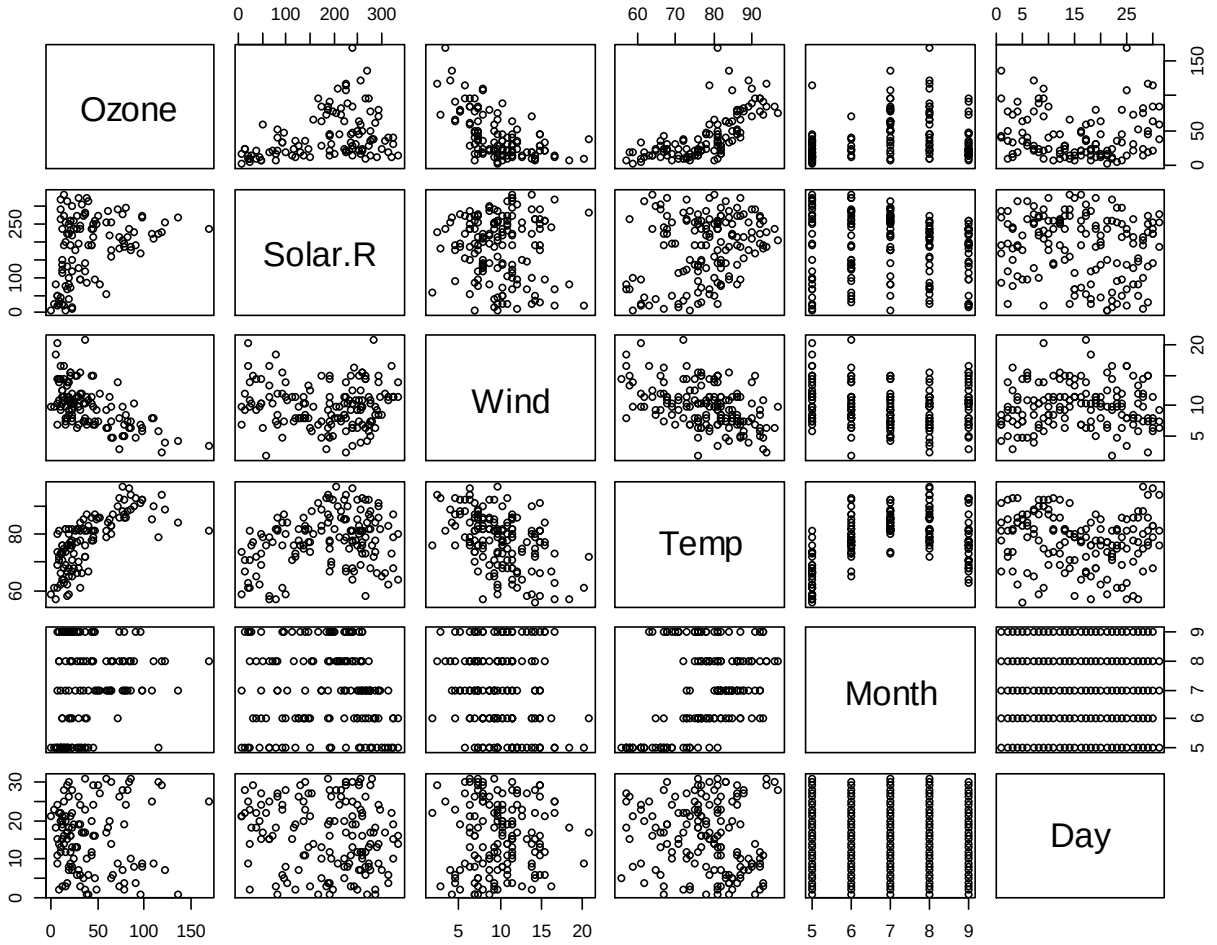


オゾンの量は日射・風・気温で説明できるか？

データを眺めるpairs()

```
pairs(airquality)
```

```
#Month・dayの情報が要らないなら
pairs(airquality[,1:4])
```



データフレームの操作

- Ozone・Solar.Rに関しては欠損値（NA）が入っている。

（このままでも解析できるが、）NAを除いたデータフレームを再定義したい

subset()関数：論理式に合うデータのみを抽出できる

```
data3.2<- subset(airquality,is.na(Ozone)==F&is.na(Solar.R)==F)
```

is.na(ベクトル)：　ベクトルの各要素に対しNAであるかどうか
NAであればTRUE、NAでなければFALSEを返す
＝＝　：　左辺の条件に右辺が合致するものを返す論理記号
論理式A＆論理式B：　論理式Aかつ論理式B

（利用例）

- 処理Aのデータだけ　`subset(data, trt==“A”)`
- 深さ10m以上のデータだけ　`subset(data, depth>=10)`
- 地点C以外のデータだけ　`subset(data, site!=“C”)`

モデルの記述

$$\mathrm{Ozone}_i = \beta_0 + \beta_1 \times \mathrm{Solar.R}_i + \beta_2 \times \mathrm{Wind}_i + \beta_3 \times \mathrm{Temp}_i + \varepsilon_i$$

切片　　偏回帰係数

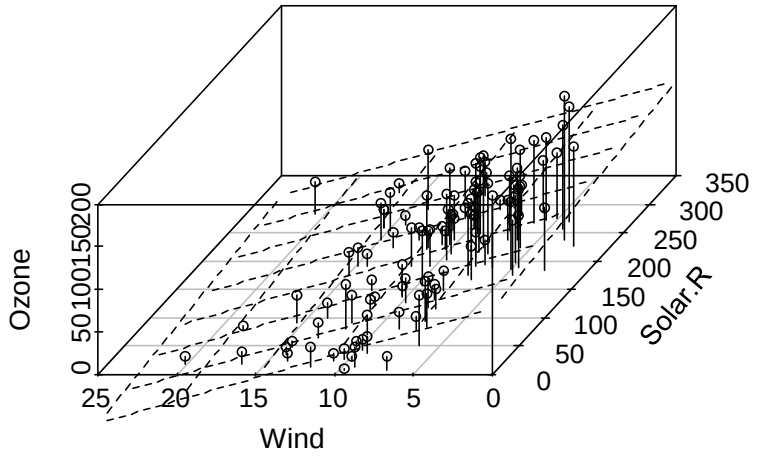
偏回帰係数：他の説明変数を固定した時のその説明変数と目的変数の関係

≠　単回帰の時の回帰係数（＝他の説明変数を無視した時の関係）

誤差を最小にする偏回帰係数を推定する（最小二乗法）

説明変数2個の場合、回帰面に対する垂線方向の距離が誤差

※説明変数3個以上だと図示不可



Rによる推定

```
> model3.2<- lm(Ozone~Solar.R+Wind+Temp,data3.2)
> summary(model3.2)

Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp, data = data3.2)

Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Temp          1.65209    0.25353   6.516 2.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.18 on 107 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

係数表

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Temp          1.65209    0.25353   6.516 2.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\mathrm{Ozone} = -64.34208 + 0.05982 \times \mathrm{Solar.R} - 3.33359 \times \mathrm{Wind} + 1.65209 \times \mathrm{Temp}$$

の関係にあると推定された

- 偏回帰係数の大小は影響力（説明力）の強さとは関係ない
  （風速をmile/hからm/sにするだけで変わる）

ちなみに単回帰で推定される回帰係数は、

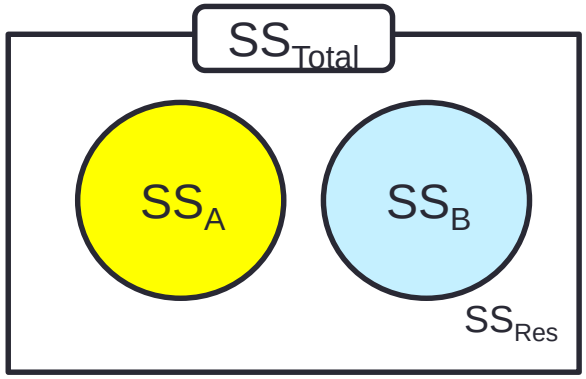
Solar.R：0.12717　Wind：-5.7288　Temp：2.4391　で偏回帰係数とは異なる

分散分析

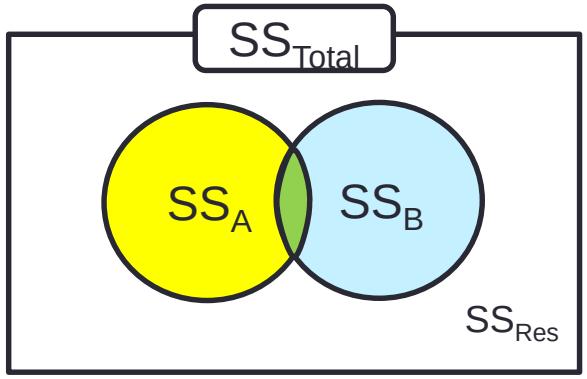
```
> anova(model3.2)
Analysis of Variance Table

Response: Ozone
           Df Sum Sq Mean Sq F value    Pr(>F)
Solar.R     1  14780   14780  32.944 8.946e-08 ***
Wind        1  39969   39969  89.094 9.509e-16 ***
Temp        1  19050   19050  42.463 2.424e-09 ***
Residuals 107  48003     449
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

SS_{Total}を各要因の平方和と残差平方和に分割

RのanovaはType I ANOVA

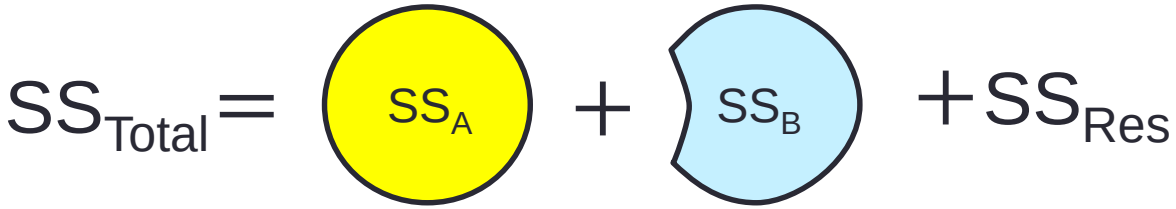


ならよいが、



実際は説明できる部分は重なる、
(多少の相関があるため)

- Rのanova関数では、説明変数をいれた順番に算出した平方和(逐次平方和)を元にしたType I ANOVAを行っている



実際に順序を入れ替えると・・・

```
Response: Ozone
           Df Sum Sq Mean Sq F value    Pr(>F)
Solar.R     1  14780   14780  32.944 8.946e-08 ***
Wind        1  39969   39969  89.094 9.509e-16 ***
Temp        1  19050   19050  42.463 2.424e-09 ***
Residuals 107  48003     449
```

Solar.R+Wind+Temp

```
Response: Ozone
           Df Sum Sq Mean Sq  F value    Pr(>F)
Temp        1  59434   59434 132.4816 < 2.2e-16 ***
Solar.R     1   2723    2723   6.0698   0.01534 *
Wind        1  11642   11642  25.9495 1.516e-06 ***
Residuals 107  48003     449
```

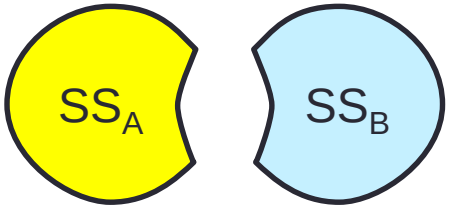
Temp+Solar.R+Wind

```
Response: Ozone
           Df Sum Sq Mean Sq  F value    Pr(>F)
Wind        1  45694   45694 101.8541 < 2.2e-16 ***
Temp        1  25119   25119  55.9905 2.149e-11 ***
Solar.R     1   2986    2986   6.6563   0.01124 *
Residuals 107  48003     449
```

Wind+Temp+Solar.R

Type II ANOVAを行うAnova()

- 順序によらない調整平方和を用いたType II ANOVAを行いたい



共通部分を差し引いた平方和を元に分散分析したい

```
> library(car)
> Anova(model3.2)
Anova Table (Type II tests)

Response: Ozone
          Sum Sq  Df F value    Pr(>F)
Solar.R     2986   1  6.6563   0.01124 *
Wind       11642   1 25.9495 1.516e-06 ***
Temp       19050   1 42.4630 2.424e-09 ***
Residuals  48003 107
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

Anova()はパッケージcarに入っている関数

ここでのSum Sqは他の変数で説明されない部分の平方和

多重共線性Multicollinearity

- 説明変数間の相関が強い（多重共線性がある）とき、重回帰の推定や解釈に問題が生じる

（例） 右手の長さ（X1）と左手の長さ（X2）で身長を推定する
単子葉被度と双子葉被度で草地の昆虫の個体数を推定する

独立で説明できる部分はほとんどないのでType II ANOVAにおいて
有意差は検出されない
偏回帰係数の推定も信頼出来ない結果に

どのようにチェック？

```
vif(model3.2)
```
carパッケージ内の関数

VIF（variance inflation factor）が大きい（＞10）と問題

```
pairs(data3.2[,2:4])
```
見た目で確認

```
cor(data3.2[,2:4])
```
相関行列で説明変数間で高い相関（|r|>0.7）がないか確認

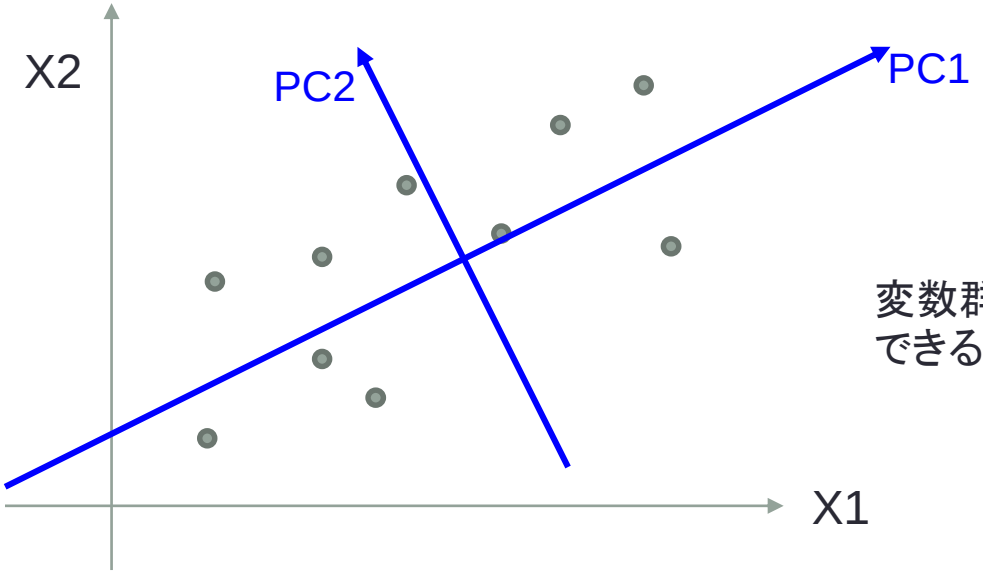
多重共線性の対処法

- 強い相関のある変数を説明変数から除く

  解釈する場合、抜いた変数が間接的に聞いている可能性も検討が必要

- 関連する変数群を合成変数にまとめる

  PCA（主成分分析）によって、複数の変数をまとめることができる
  ※詳しい方法は多変量解析のときにでも・・・



変数群の持つばらつきをうまく表現できる新たな軸を設定してやる

重回帰における交互作用

- 二元配置分散分析と同様に重回帰においても交互作用を設定できる

Solar.RとWindの交互作用を想定

```
> model3.2i<- lm(Ozone~Solar.R*Wind,data3.2)
> summary(model3.2i)

Call:
lm(formula = Ozone ~ Solar.R * Wind, data = data3.2)

Residuals:
    Min      1Q  Median      3Q     Max
-48.694 -17.200  -4.384  12.740  78.218

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.467686  17.634602   1.955 0.053246 .
Solar.R       0.324141   0.083928   3.862 0.000193 ***
Wind         -1.594546   1.508979  -1.057 0.293026
Solar.R:Wind -0.020279   0.007246  -2.799 0.006089 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.16 on 107 degrees of freedom
Multiple R-squared:  0.487,     Adjusted R-squared:  0.4727
F-statistic: 33.86 on 3 and 107 DF,  p-value: 1.807e-15
```

重回帰における交互作用

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.467686  17.634602   1.955 0.053246 .
Solar.R       0.324141   0.083928   3.862 0.000193 ***
Wind         -1.594546   1.508979  -1.057 0.293026
Solar.R:Wind -0.020279   0.007246  -2.799 0.006089 **
```

$$\text{Ozone} = 34.5 + 0.032 \times \text{Solar.R} - 1.59 \times \text{Wind} - 0.02 \times \text{Solar.R} \times \text{Wind}$$

交互作用は説明変数同士を
掛けあわせたものに対する係数

```
Anova Table (Type II tests)

Response: Ozone
             Sum Sq  Df F value    Pr(>F)
Solar.R        9055   1 15.5072 0.0001467 ***
Wind          39969   1 68.4503  3.93e-13 ***
Solar.R:Wind   4573   1  7.8321 0.0060892 **
Residuals     62479 107
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

あれ？Wind有意

係数表では偏回帰係数は有意じゃないのに

交互作用に伴う問題

- X_1の偏回帰係数はX_2がゼロの時の効果を見ている･･･本当に意味ある？

（例） 身長と体重とその交互作用から50m走の速さを推定
体重0kgの時の身長の効果が偏回帰係数として得られる

- X_1とX_2に相関がなくても、X_1とX_1:X_2やX_2とX_1:X_2に相関が生じて、多重共線性の問題が生じる

対処するには、説明変数を**センタリング（中央化）や標準化する**

平均をゼロにする
（＝データを平均値で引く）

平均ゼロ･標準偏差1にする
（＝平均値で引いてSDで割る）

```
data3.2$Solar.Rc<- scale(data3.2$Solar.R,scale=F)
data3.2$Windc<- scale(data3.2$Wind,scale=F)
```

#各変数をセンタリング
#scale=Tで標準化

```
model3.2ic<- lm(Ozone~Solar.Rc*Windc,data3.2)
```

#センタリングした変数で再計算

センタリングした変数での重回帰

```
> model3.2ic<- lm(Ozone~Solar.Rc*Windc,data3.2)
> summary(model3.2ic)

Call:
lm(formula = Ozone ~ Solar.Rc * Windc, data = data3.2)

Residuals:
    Min      1Q  Median      3Q     Max
-48.694 -17.200  -4.384  12.740  78.218

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.270223   2.312630  17.846  < 2e-16 ***
Solar.Rc        0.122573   0.026692   4.592 1.20e-05 ***
Windc          -5.342176   0.653253  -8.178 6.41e-13 ***
Solar.Rc:Windc -0.020279   0.007246  -2.799  0.00609 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.16 on 107 degrees of freedom
Multiple R-squared:  0.487,     Adjusted R-squared:  0.4727
F-statistic: 33.86 on 3 and 107 DF,  p-value: 1.807e-15
```

Anovalによる検定結果も概ね一致
（多少の共線性あるのでずれる）

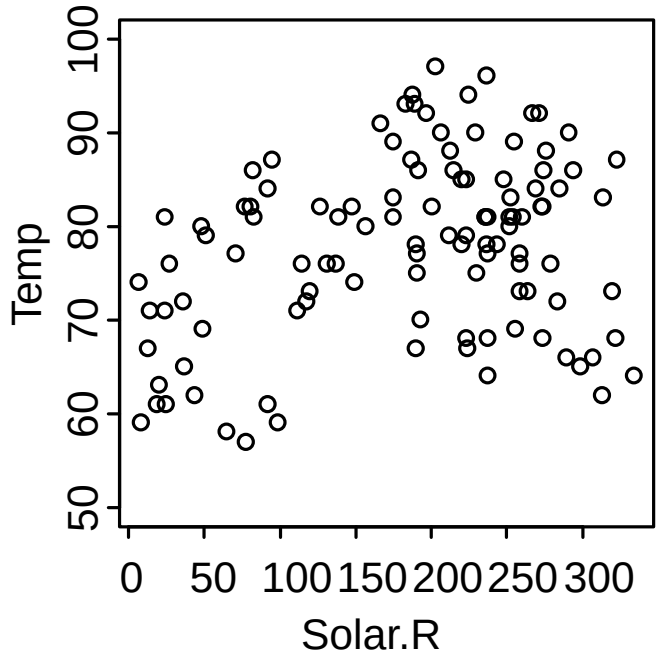
```
Anova Table (Type II tests)

Response: Ozone
              Sum Sq  Df F value    Pr(>F)
Solar.Rc        9055   1 15.5072 0.0001467 ***
Windc          39969   1 68.4503  3.93e-13 ***
Solar.Rc:Windc  4573   1  7.8321 0.0060892 **
Residuals      62479 107
---
```

$$\mathrm{Ozone} = 41.3 + 0.12 \times \left(\mathrm{Solar.R} - \overline{\mathrm{Solar.R}}\right) - 5.34 \times \left(\mathrm{Wind} - \overline{\mathrm{Wind}}\right)$$
$$-0.02 \times \left(\mathrm{Solar.R} - \overline{\mathrm{Solar.R}}\right) \times \left(\mathrm{Wind} - \overline{\mathrm{Wind}}\right)$$

偏回帰係数は他の説明変数が平均値の時の効果を表す（解釈しやすい）

二次回帰

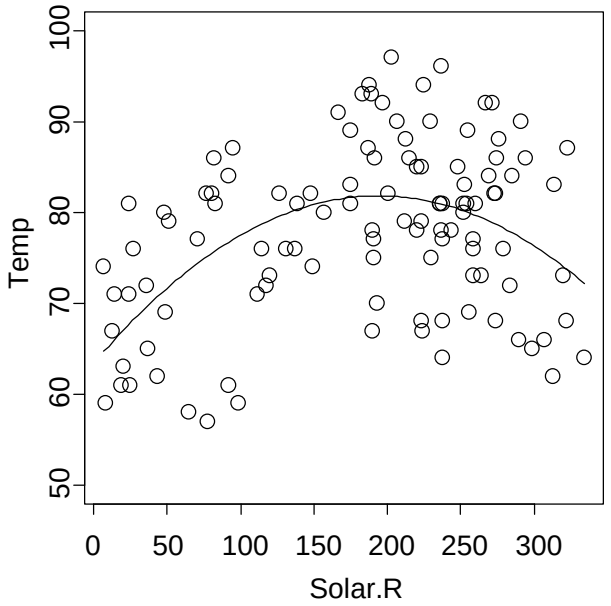
- 一山形（もしくは下に凸）の関係性を記述したい場合、二次項（X^2）を説明変数に加える事で記述できる $y=ax^2+bx+c$にあてはめる
- この場合も多重共線性を防ぐためにセンタリングしたものの二乗を用いる



一山形？頭打ち？
いずれにしろ直線ではなさそう

二次回帰による推定

```
> data3.2$Solar.Rc2<- data3.2$Solar.Rc^2                #二乗項
> model3.2sq<- lm(Temp~Solar.Rc+Solar.Rc2,data3.2)      #二次回帰モデル
> summary(model3.2sq)

Call:
lm(formula = Temp ~ Solar.Rc + Solar.Rc2, data = data3.2)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4451  -5.8855   0.1407   7.0106  15.1564

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.8417453  1.1812131  69.286  < 2e-16 ***
Solar.Rc     0.0090500  0.0099457   0.910    0.365
Solar.Rc2   -0.0004917  0.0001059  -4.643 9.74e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.393 on 108 degrees of freedom
Multiple R-squared:  0.2385,    Adjusted R-squared:  0.2244
F-statistic: 16.91 on 2 and 108 DF,  p-value: 4.086e-07
```



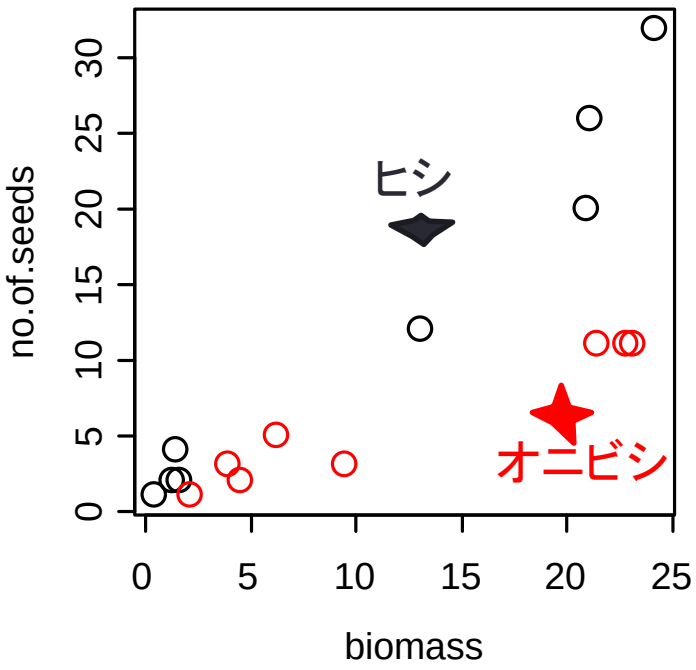
$$\mathrm{Temp} = 81.8 + 0.009 \times \left(\mathrm{Solar.R} - \overline{\mathrm{Solar.R}}\right) - 0.00049 \times \left(\mathrm{Solar.R} - \overline{\mathrm{Solar.R}}\right)^2$$

```
plot(Temp~Solar.R,data3.2,cex=2,ylim=c(50,100))
curve(81.8+0.009*(x-mean(data3.2$Solar.R))-0.00049*(x-mean(data3.2$Solar.R))^2,add=T)
```

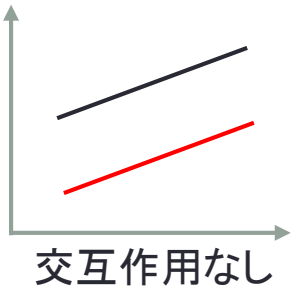
共分散分析ANCOVA

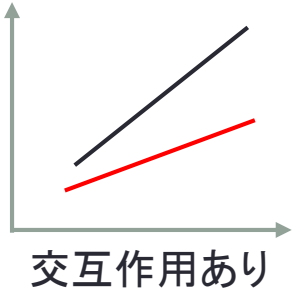
- 説明変数に連続変数とカテゴリカル変数が混在する場合も同様な解析を行うことができる

```
data3.1<- read.csv("data3.1.csv",T)
plot(no.of.seeds~biomass,col=species,cex=2,data3.1)
```



- バイオマスを考慮するとヒシとオニビシの種子数に差が見られるのでは？
- ヒシとオニビシではバイオマスと種子数の関係性に違いがあるのでは？（種類とバイオマスの交互作用）





Rによる推定

```
> data3.1$biomassc<- scale(data3.1$biomass,scale=F)
> model3.3<- lm(no.of.seeds~biomassc*species,data3.1)
> summary(model3.3)

Call:
lm(formula = no.of.seeds ~ biomassc * species, data = data3.1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2761 -0.4539  0.0374  0.9593  4.0045

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           13.06900    0.78697  16.607 1.21e-09 ***
biomassc               1.14126    0.08049  14.179 7.38e-09 ***
speciesnatans         -7.47683    1.11326  -6.716 2.15e-05 ***
biomassc:speciesnatans -0.67616    0.12193  -5.546 0.000127 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.222 on 12 degrees of freedom
Multiple R-squared:  0.9561,    Adjusted R-squared:  0.9451
F-statistic: 87.03 on 3 and 12 DF,  p-value: 2.071e-08
```

#バイオマスをセンタリング
#交互作用込みでモデリング

推定値の解釈

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             13.06900    0.78697  16.607 1.21e-09 ***
biomassc                 1.14126    0.08049  14.179 7.38e-09 ***
speciesnatans           -7.47683    1.11326  -6.716 2.15e-05 ***
biomassc:speciesnatans -0.67616    0.12193  -5.546 0.000127 ***
---
```

speciesnatans・・・ speciesのカテゴリーnatansがjaponica(基準)に比べてどうか
基準となるカテゴリーはアルファベット順で最初のもの
センタリングしてあるので、バイオマスが平均値の時の関係性

biomassc:speciesnatans
・・・ speciesのカテゴリーnatansにおけるbiomassの傾きが
japonica(基準)におけるbiomassの傾きに比べてどうか

japonicaに関しては

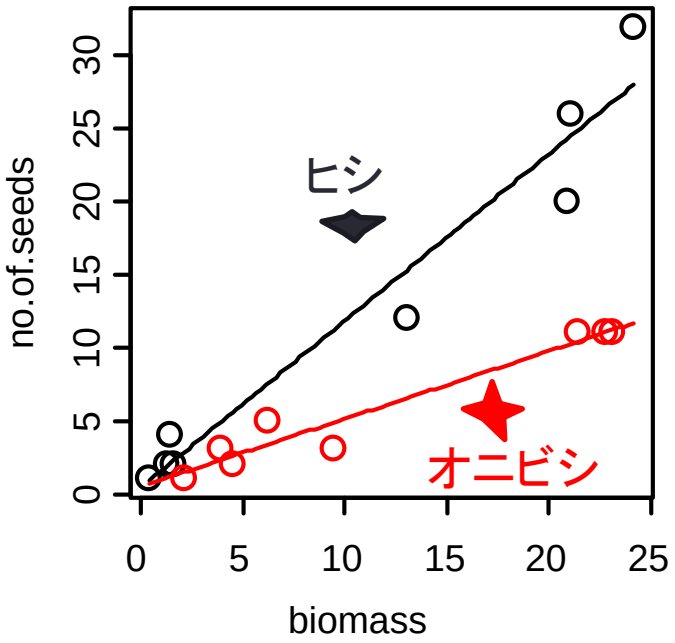
$$\mathrm{no.\,of.\,seeds} = 13.1 + 1.14 \times \left(\mathrm{biomass} - \overline{\mathrm{biomass}}\right)$$

natansに関しては

$$\mathrm{no.\,of.\,seeds} = 13.1 - 7.48 + (1.14 - 0.676) \times \left(\mathrm{biomass} - \overline{\mathrm{biomass}}\right)$$

ANCOVAの結果の図示

```
plot(no.of.seeds~biomass,col=species,cex=2,data3.1)
curve(13.1+1.14*(x-mean(data3.1$biomass)),add=T)
curve(13.1-7.48+(1.14-0.676)*(x-mean(data3.1$biomass)),add=T,col=2)
```



- 主効果でspeciesが有意

→同じバイオマス（平均値）で比べると、ヒシはオニビシに比べ種子数が多い

- 交互作用が有意

→ヒシのほうがバイオマスの増加に対する種子数の増加率が高い

まとめ：一般線形モデル

- $y_{ij} = \beta_0 + \sum \beta_i \times x_{ij} + \varepsilon_{ij}$ の形で記述でき、

  Rでは lm（Y~X1+X2＋・・・）の形で記述し、推定する

- 説明変数のタイプによって分散分析・重回帰分析・共分散分析などに分けられるが、同じ枠組みで扱える
- 交互作用の考慮や二次項を用いることで、非相加的関係や非線形関係も推定できる

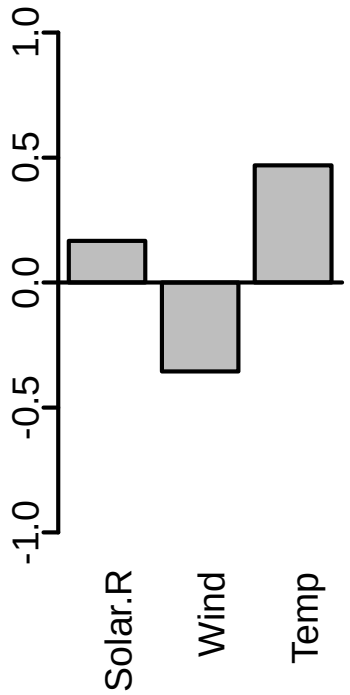
- 分散分析のタイプや変数のセンタリングによって結果や解釈が変わってくるので注意
- 説明変数間での偏回帰係数の大小は影響の強さとは関係ない
- 誤差の正規分布・等分散を仮定しているので、前提が満たされないデータの当てはめには不向き（→他の分布を仮定：一般化線形モデル）

偏回帰係数の比較

- もとの変数を標準化（平均0、標準偏差1）に変換しておけば、偏回帰係数間の比較が可能（標準化偏回帰係数）

```
> data3.2s<- data.frame(scale(data3.2))
> model3.2s<- lm(Ozone~Solar.R+Wind+Temp,data3.2s)
> summary(model3.2s)

Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp, data = data3.2s)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2166 -0.4273 -0.1067  0.3034  2.8735

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.815e-17  6.042e-02   0.000   1.0000
Solar.R      1.639e-01  6.351e-02   2.580   0.0112 *
Wind        -3.564e-01  6.997e-02  -5.094 1.52e-06 ***
Temp         4.731e-01  7.261e-02   6.516 2.42e-09 ***
---(略)
```

#データフレームの各列をまとめて標準化
#標準化データで重回帰



気温、風、日射の順に影響が強そう

```
barplot(coefficients(model3.2s)[-1],ylim=c(-1,1),las=3)
abline(h=0)
```

正規性のチェック

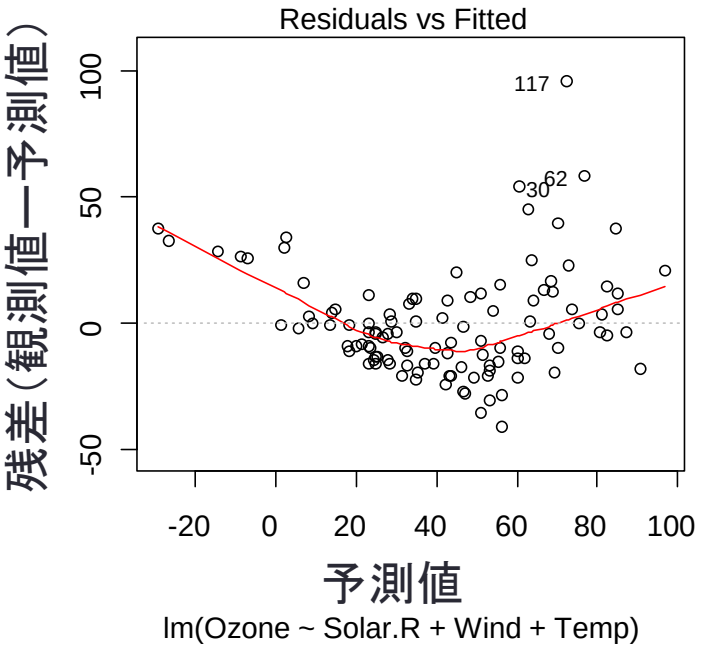
$$y_{ij} = \beta_0 + \sum \beta_i \times x_{ij} + \varepsilon_{ij}$$

- 誤差ε_{ij}が正規分布・等分散かどうかをチェックするには？

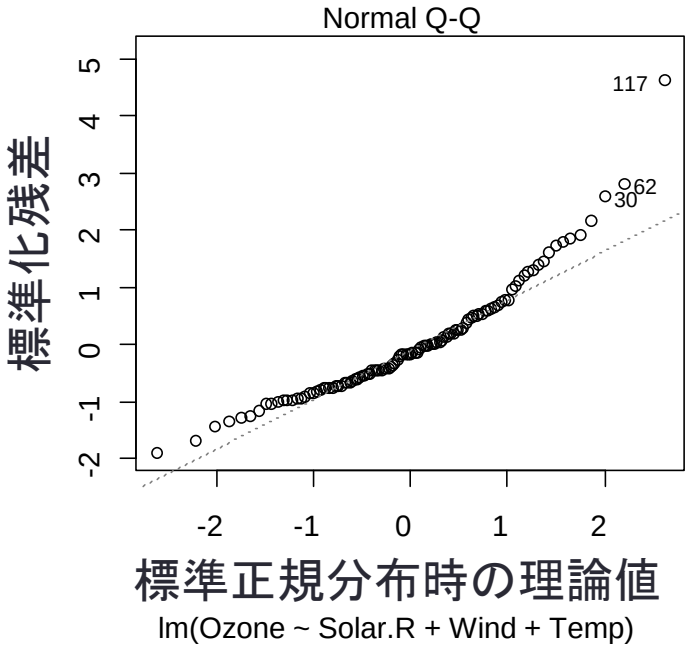
plot(モデル)で確認可能

```
plot(model3.2)
```

等分散ならばらつきが予測値に対し一様



正規分布しているなら直線に乗る



→若干問題あり？
変数変換したほうがよいかも

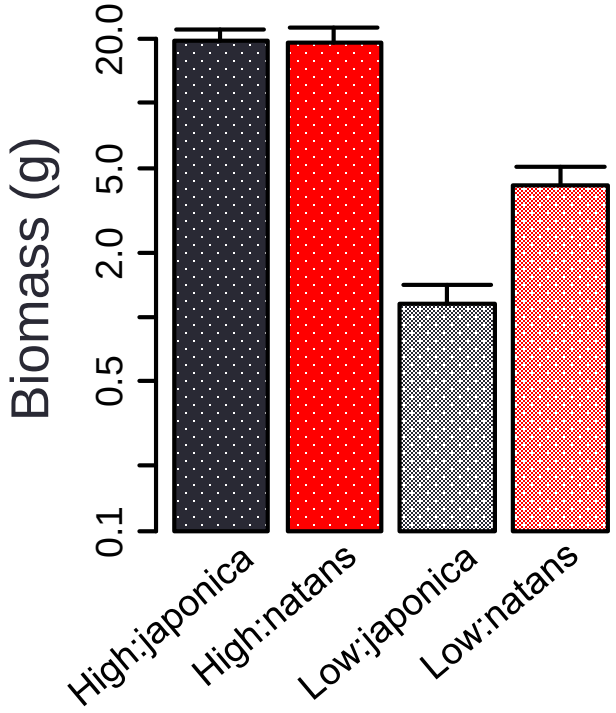
その他一般化加法モデル（GAM）など別のモデルに当てはめる

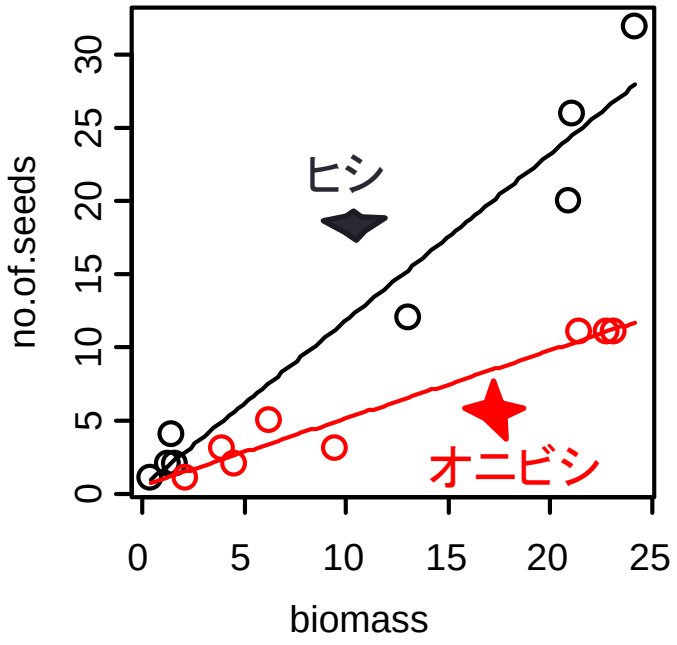
結果の表現

一般線形モデルのグラフィカルな表現方法

一般線形モデルの結果

- 2元配置分散分析（棒グラフ）やANCOVA（散布図＋回帰線）に関しては、前回の作図の応用で対処可能
- 解析結果は、係数表や分散分析表で示す

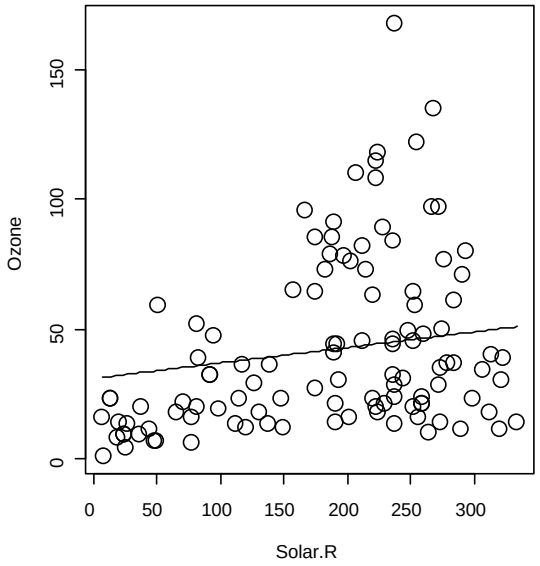


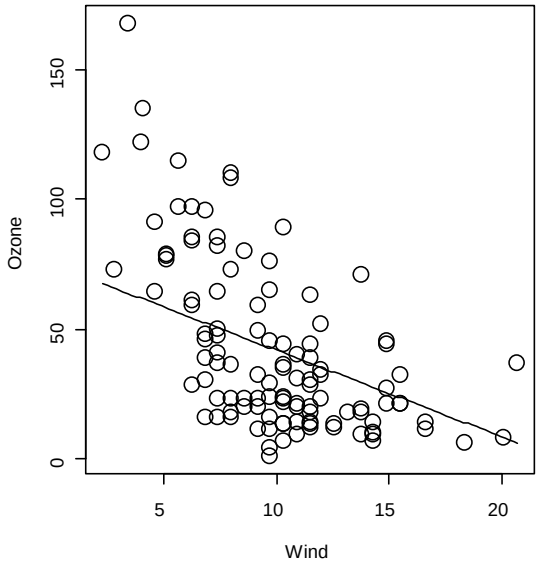


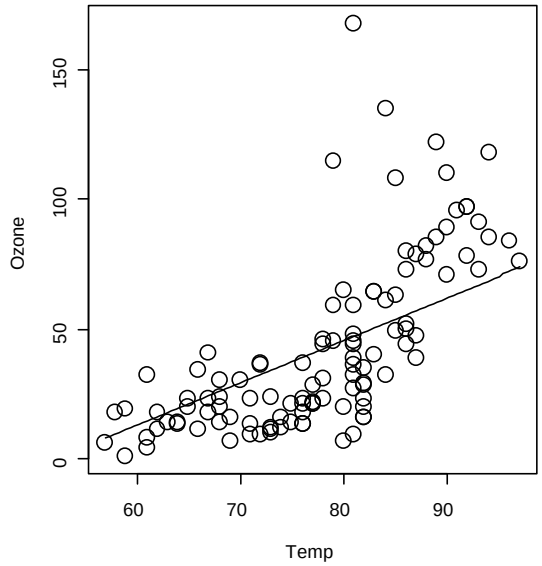
重回帰分析の表現

- 散布図に、重回帰分析で得られる偏回帰係数をもとにした回帰線を引いてもあまり意味がない（特に交互作用があったり変数が多い場合）

（他の変数で説明されるばらつきが大きい場合、見かけの傾向と直線が合わなかったり、極端な場合は、みかけの関係性と逆になる）







※　この程度ならそれほど違和感ないかも

どのように表現するか

- 論文中では、係数表と分散分析表だけでも十分（各変数の偏回帰係数の推定値及びSEをのせる）
- プレゼンなどでグラフィカルに表現したい場合

#観測値と説明変数の傾向を見せたい場合

- 回帰線書かないで散布図のみ

#観測値と重回帰の関係性を見せたい場合

- 説明変数2つまでなら3次元プロットと回帰面をのせる ★
- 注釈つけた上で偏回帰係数による回帰線をのせる

#一山形・頭打ちなど回帰による予測線の形状を見せたい場合

- 散布図無しで偏回帰係数による回帰線をのせる

#回帰によるばらつきの説明具合を見せたい場合

- 見たい変数以外の重回帰による残差をもとにした、偏回帰プロットをのせる ★

#説明変数間での影響の強さの比較を見せたい場合

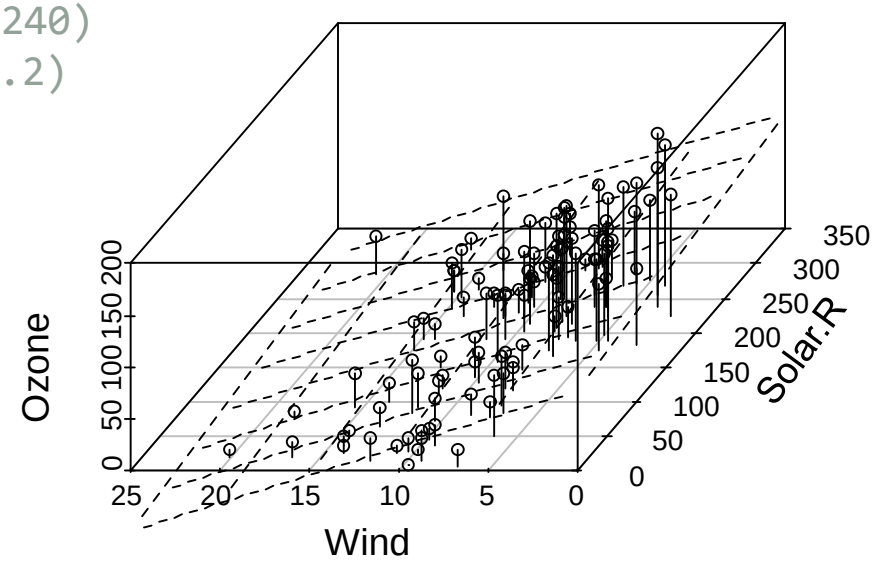
- 標準化偏回帰係数または偏回帰係数のEstimate/SEを棒グラフで示す ★

3次元プロット

- 目的変数と2種の説明変数の関係性について3次元的に表現したもの
- いろいろ角度等を調整しないと結構見づらい場合も

library(scatterplot3d)内のscatterplot3d関数を用いる

```
library(scatterplot3d)
s3d<- scatterplot3d(x=data3.2$Solar.R,y=data3.2$Wind,z=data3.2$Ozone,
                    type="h",angle=240)
mod3d<- lm(Ozone~Solar.R+Wind,data3.2)
s3d$plane3d(mod3d)
```

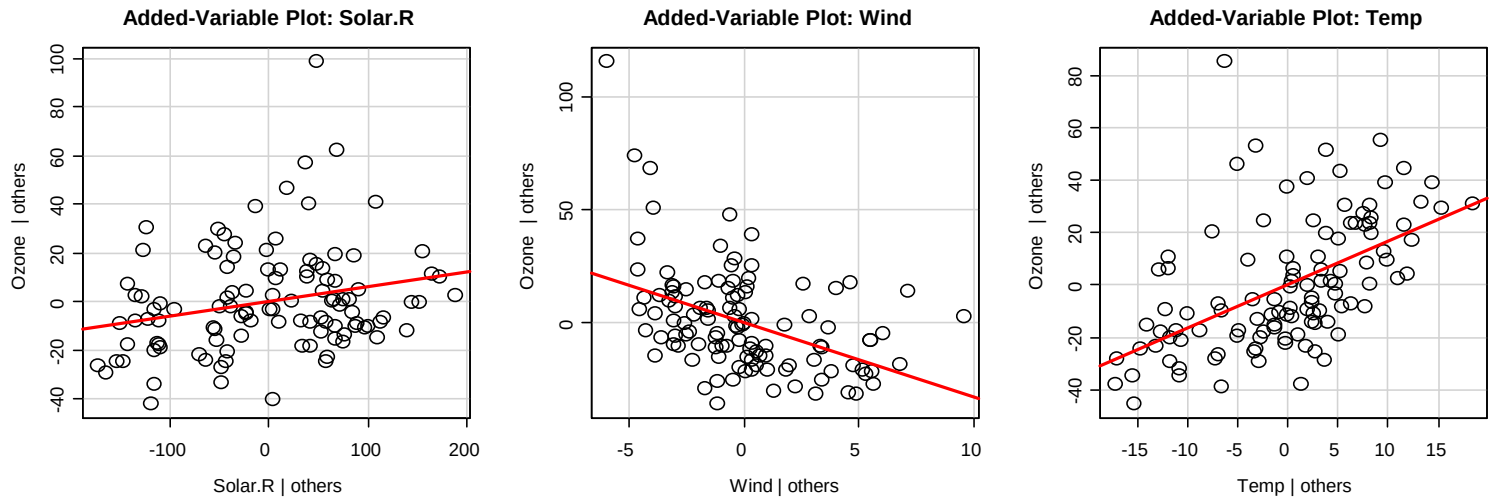


偏回帰プロット

- 縦軸は見たい変数X_i以外の変数でYを重回帰した時の残差
- 横軸は見たい変数X_i以外の変数でX_iを重回帰した時の残差
- 回帰線の傾きは偏回帰係数に一致

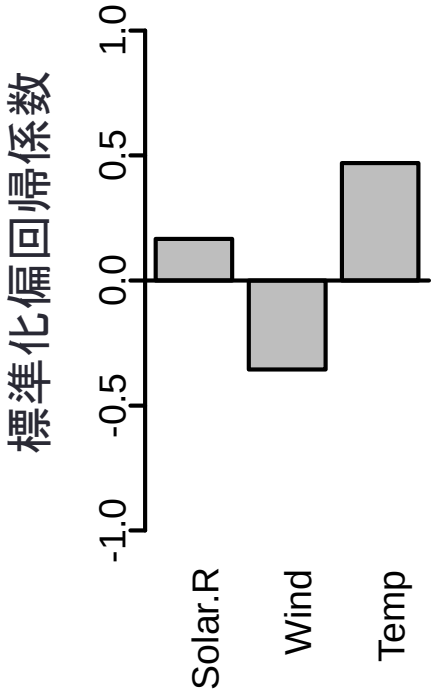
```
model3.2<- lm(Ozone~Solar.R+Wind+Temp,data3.2)
par(mfrow=c(1,3))
avPlot(model3.2,"Solar.R",cex=2)
avPlot(model3.2,"Wind",cex=2)
avPlot(model3.2,"Temp",cex=2)
```

```
#par(mfrow=c(row,column))で描画領域を分割
#library(car)のavPlot関数で書ける
```

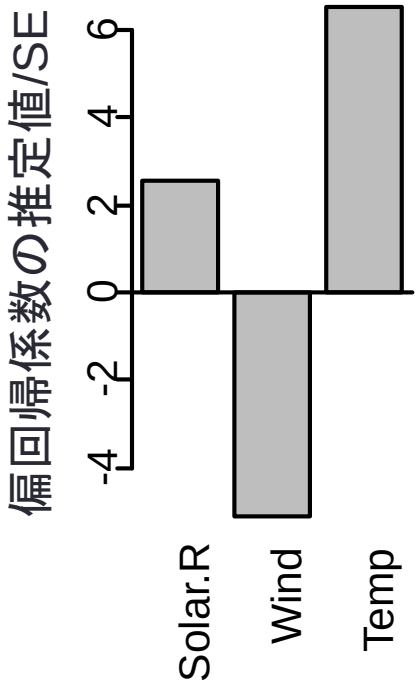


偏回帰係数のグラフィカルな比較

偏回帰係数を比較可能な形で棒グラフで図示することで、変数間の相対的な影響力を比較




```
barplot(coefficients(model3.2s)[-1],ylim=c(-1,1),las=3)
abline(h=0)
```




```
barplot(summary(model3.2)$coefficients[-1,3],las=3)
abline(h=0)
```

およそ±2を超えるとt検定において(5%水準で)有意になるのでわかりやすい

ただ、統計の本でこの表現を見たことないので、もしかしたら統計的には問題のある比較なのかも
(要確認)

次回予告

- R編：未定（なし？）

- 統計編：一般化線形モデル（GLM）
  - 最尤法
  - 0/1データの解析（二項分布）
  - カウントデータの解析（ポアソン分布・負の二項分布）
  - 赤池情報量基準AICとモデル選択

- 表現編：一般化線形モデルの結果