Rを利用した回帰分析

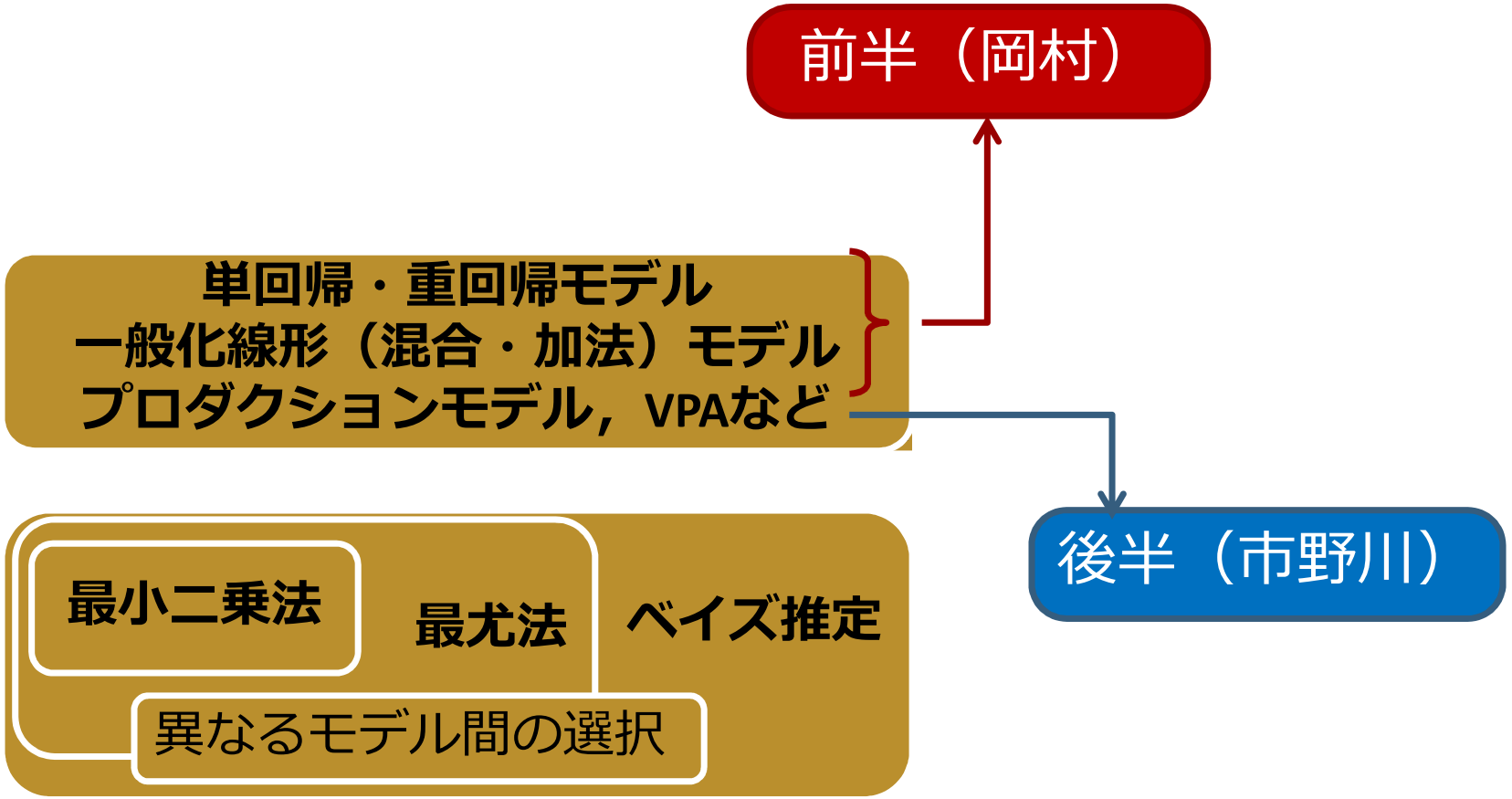
中央水産研究所
岡村　寛

水産資源学における統計解析

- 漁業・調査データ解析
  - CPUE標準化 ～ 資源のトレンド
  - 体長組成のモード分解
- 成長式などの生物パラメータの推定
- 資源評価モデルによる個体群評価

→　**ほとんどがパラメータの推定問題**

今日の概要


前半（岡村）
単回帰・重回帰モデル
一般化線形（混合・加法）モデル
プロダクションモデル，VPAなど
後半（市野川）
最小二乗法
最尤法
ベイズ推定
異なるモデル間の選択

研修の成功と失敗

・R初心者

成功☺: 自分にもできそう，面白そう，仕事に役立ちそう

失敗☹: 自分にはできないな，今までどおりExcelで…

・R経験者

成功☺: そういうときのプログラムはこう書くのか，こんなパッケージがあるのか

失敗☹: 全部知ってることでつまらない，私ならもっとうまく…

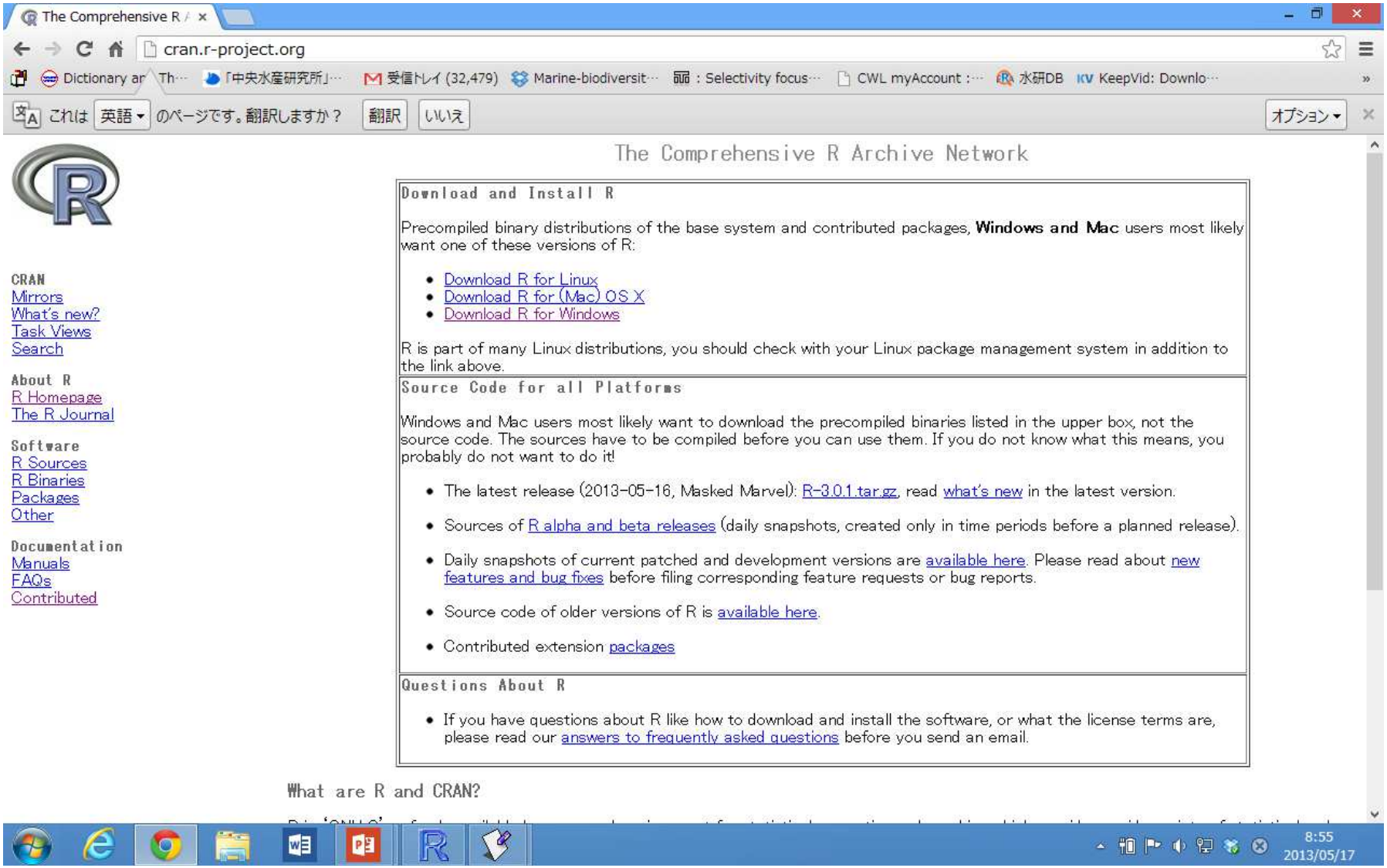
Why R?

- 無料！
- 既存の統計処理をほぼ網羅
- 組み合わせて新しい解析を
- 乱数発生・シミュレーションが容易
- 気軽にプログラミング
- グラフィックス
- 他の言語（WinBUGS, ADMB, …）を呼び出して使用

前半の目的

- Rを利用したデータ解析のやり方に慣れる
- 回帰/GLMの考え方を理解する
- Rによる結果の解釈
- 結果のグラフ化
- プログラミングの基礎
- GLMM/GAM/VGAMについてなんとなく理解

R


The Comprehensive R
cran.r-project.org
これは 英語 のページです。翻訳しますか？ 翻訳 いいえ オプション
The Comprehensive R Archive Network
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Download R for Linux
Download R for (Mac) OS X
Download R for Windows
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.
Source Code for all Platforms
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
The latest release (2013-05-16, Masked Marvel): R-3.0.1.tar.gz, read what's new in the latest version.
Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
Source code of older versions of R is available here.
Contributed extension packages
Questions About R
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.
What are R and CRAN?
8:55
2013/05/17

Working directory

作業するフォルダを設定してやる

- getwd()
- setwd("C:/Rkenshu")
- getwd()

- q() → save workspace image? Yesなら次回から.Rdataをダブルクリックすれば前回の作業から続けられる

データの読み込み

- scan

  scan("dat1.dat")

- read.table, read.csv, read.fwf

  bp.dat <- read.csv(“bloodpressure.csv”)

- load

  load(“mH1.rda”)

データの書き込み

- cat

  dat1 <- letters[1:10]; cat(dat1,file="dat1.dat")

- write.table, write.csv

  write.csv(bp.dat, "bp_dat.csv", row.names=FALSE)

- save

  mH1 <- lm(High~Day,data=bp.dat); save(mH1, file="modelH.rda")

データの型

```
x <- 1
class(x)
class(as.matrix(x))
class(as.data.frame(x))
class(as.integer(x))
class(as.character(x))
is.character(x)
is.numeric(x)
is.vector(x)
```

パッケージ

- library(MASS)

その他，本日使用するパッケージ

- MuMIn
- lme4
- VGAM
- mgcv

例データ

- bloodpressure.xls
- まずcsvファイルにしてやる
- Rに読み込む

  bp.dat <- read.csv(“bloodpressure.csv”)

データ概要

class(bp.dat)

names(bp.dat)

head(bp.dat)

?head

summary(bp.dat)

```
> c(mean(bp.dat$High), mean(bp.dat$Low))
[1] 136.36  90.92
```

血圧基準値

・正常血圧　125/80未満

・**高血圧　135/85以上**

男性血圧平均

年齢	
20～24：	128 / 75
25～29：	128 / 75
30～34：	129 / 77
35～39：	130 / 79
40～44：	132 / 81
45～49：	136 / 83
50～54：	144 / 87
55～59：	150 / 88
60～64：	156 / 91
65～69：	158 / 89
70以上：	165 / 89

女性血圧平均

年齢	
20～24：	121 / 72
25～29：	122 / 73
30～34：	124 / 75
35～39：	127 / 78
40～44：	132 / 80
45～49：	140 / 84
50～54：	147 / 86
55～59：	150 / 88
60～64：	158 / 90
65～69：	166 / 91
70以上：	171 / 91

教えて! goo

血圧を下げる方法を教えてください。

**薬を飲む方法以外に簡単に血圧を下げる方法を教えてください。
現在、下が９０～１１０　上は１４０～１６０　週２回水泳を１時間位やっていますが下がりません。（半年以上）
アルコールを飲んだ時に計ると８０～１３０位に下がります。
血圧を下げるのに成功した方よろしくお願いします。**

高血圧の原因

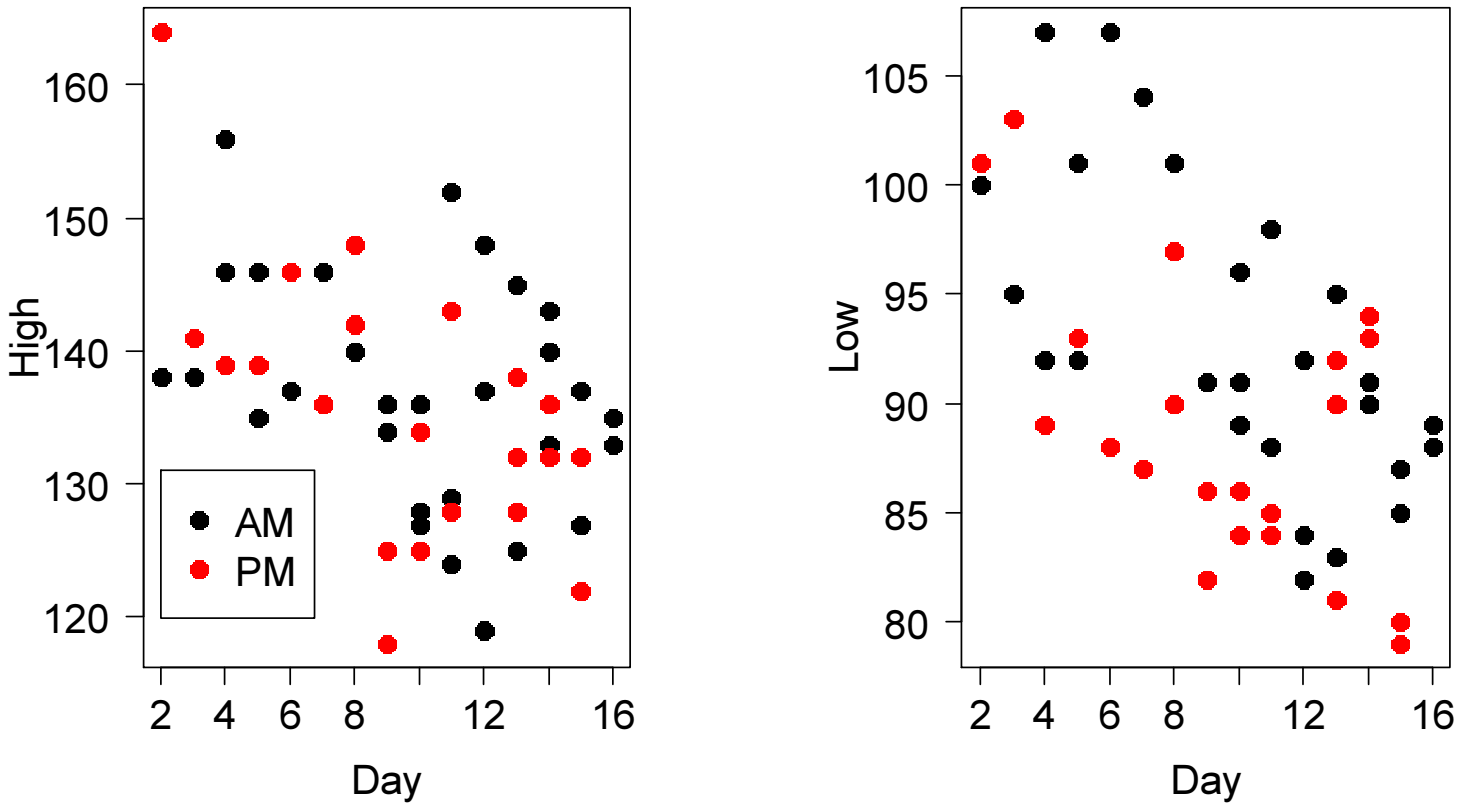
- 遺伝
- 塩分の取りすぎ
- 運動不足
- 肥満
- 加齢
- ストレス
- 気温
- 過度の飲酒と喫煙

高血圧になりやすいかチェック

- **濃い味つけのものが好き**
- 野菜や果物はあまり食べない
- **運動をあまりしない**
- 家族に高血圧の人がいる
- **ストレスがたまりやすい**
- お酒をたくさん飲む
- たばこを吸う
- 血糖値が高いといわれたことがある
- **炒めものや揚げもの、肉の脂身など、あぶらっぽい食べ ものが好き**

チェックの数が多いほど、高血圧になりやすいので、注意が必要です。

図

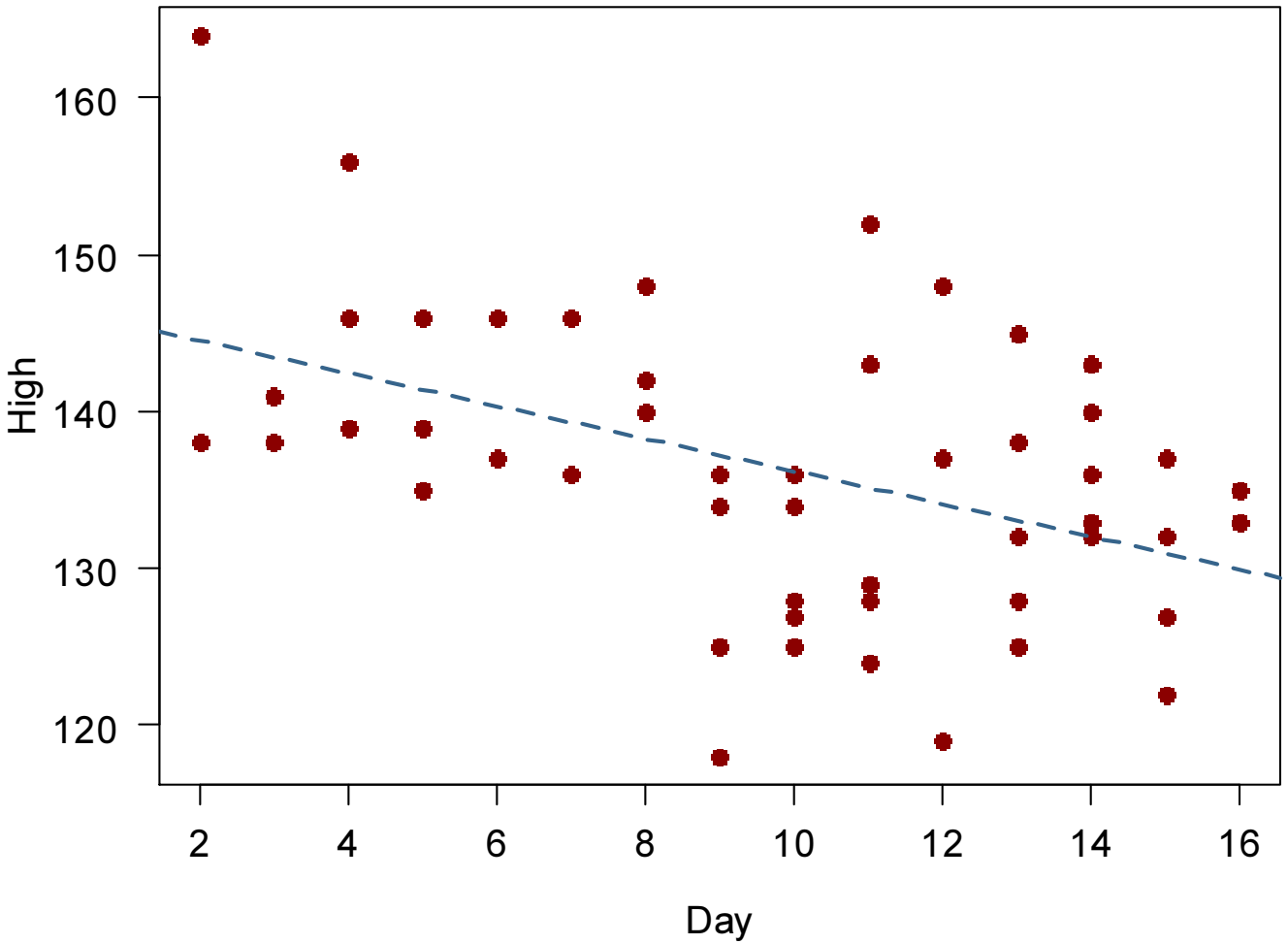

High
160
150
140
130
120
AM
PM
2
4
6
8
12
16
Day
Low
105
100
95
90
85
80
2
4
6
8
12
16
Day

回帰分析

$Y = a + bX + e, \quad e \sim N(0, \sigma^2)$

```
modelH.1 <- lm(High~Day,data=bp.dat)

summary(modelH.1)$coef

confint(modelH.1,level=0.9)
```

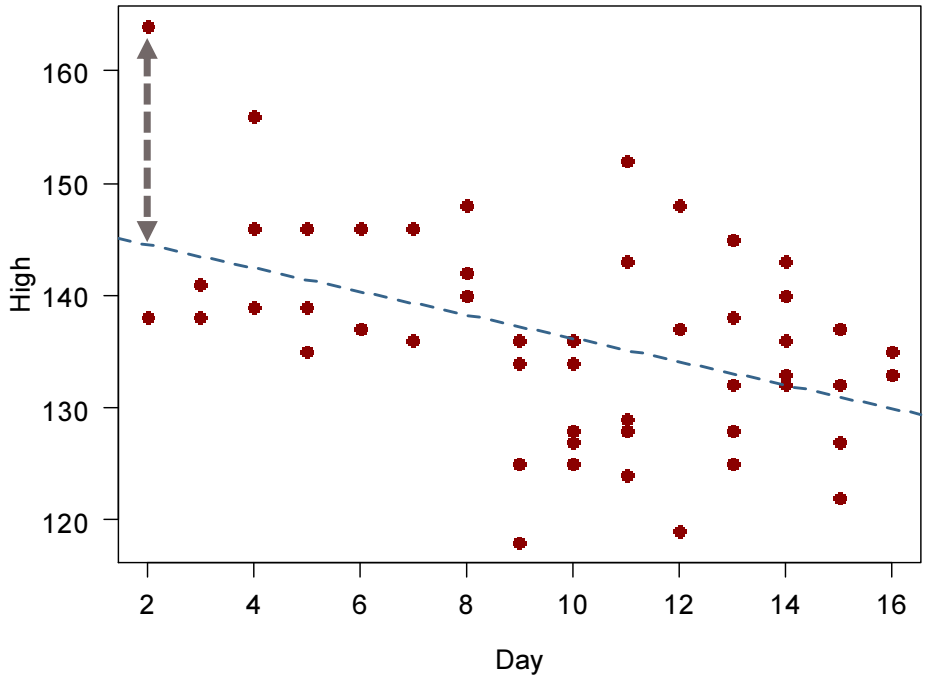
図描画



最小二乗法

$(y - (a + bx))^2$を最小化してパラメータを推定

bの推定値 = cov(x,y)/var(x)
aの推定値 = E(y) – bの推定値×E(x)



最尤推定法

$y = a + bx + e,\quad e \sim N(0, \sigma^2)$

$$\Pr(y|a,b) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y-(a+bx)\right)^2}{2\sigma^2}\right)$$

Pr(y|a,b)をa, bの関数とみなして，その関数（尤度関数）の最大化によりパラメータ推定

上の最大化は，exp(…)の中を最小にすることにより達成できる

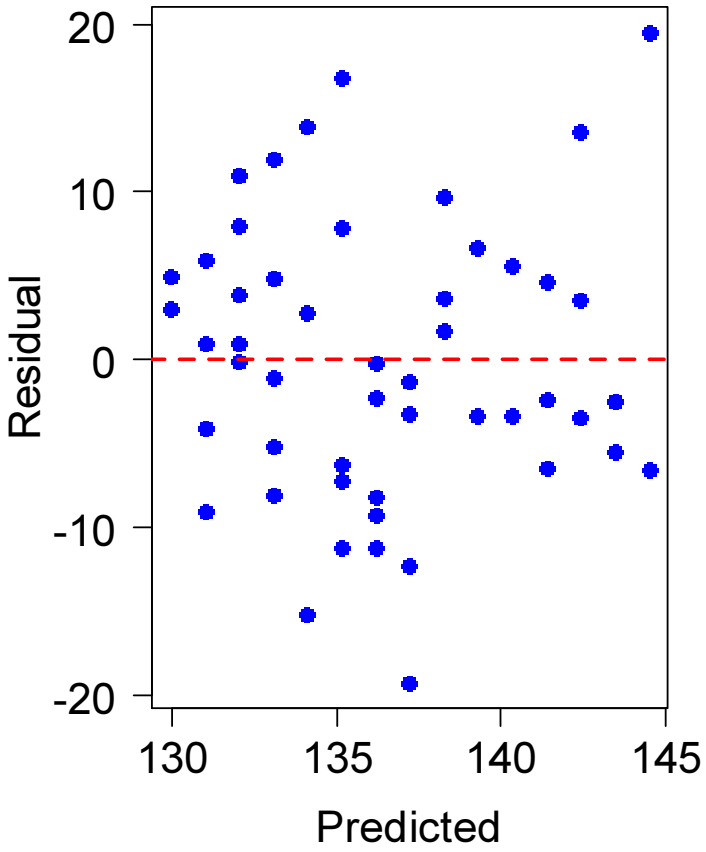
Kullback-Leibler距離

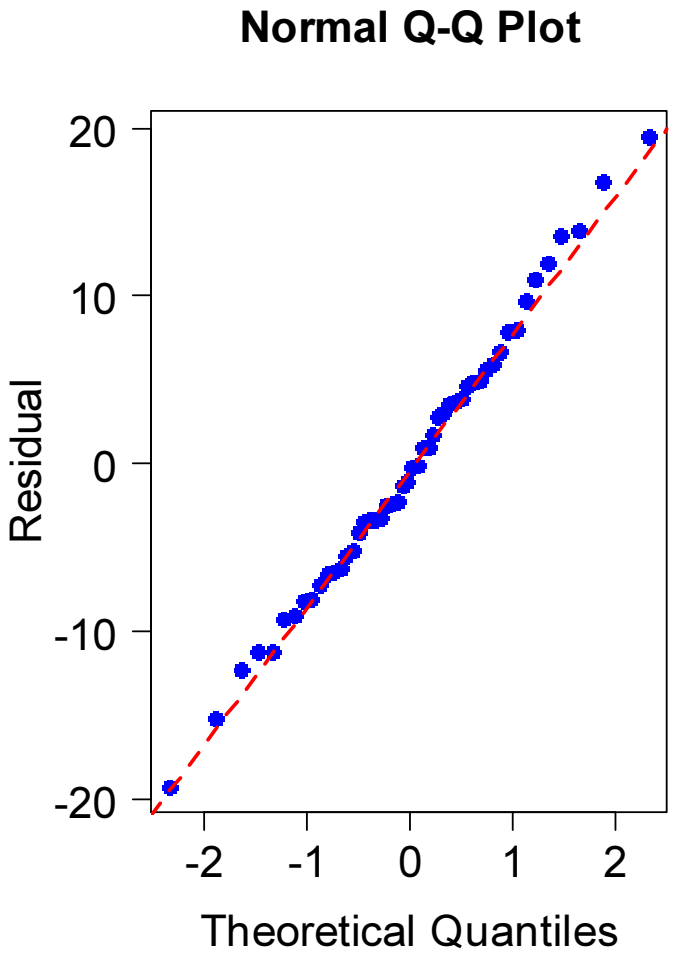
確率分布間の距離
真の分布 $f(x)$，モデル $g(x|\theta)$

Kullback-Leibler距離 $E[\log\{f(x)/g(x|\theta)\}] = E[\log(f(x)) - \log(g(x|\theta))]$
$= E[\log(f(x))] - E[\log(g(x|\theta))]$
$E[\log(g(x|\theta))] \approx (1/n) \Sigma \log(g(x|\theta))$

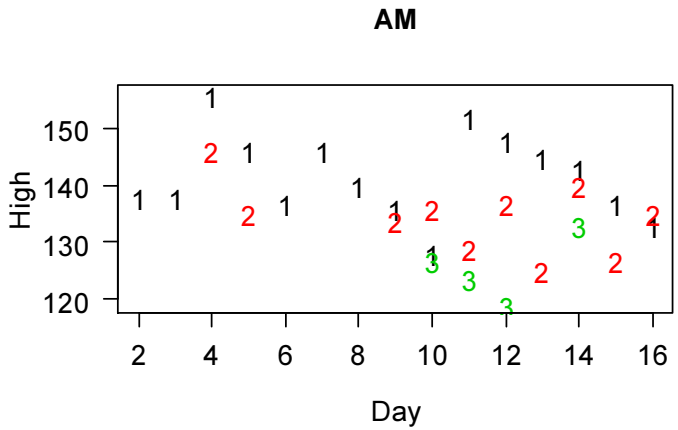
（対数）尤度を最大にするパラメータはKullback-Leibler距離を最小にする

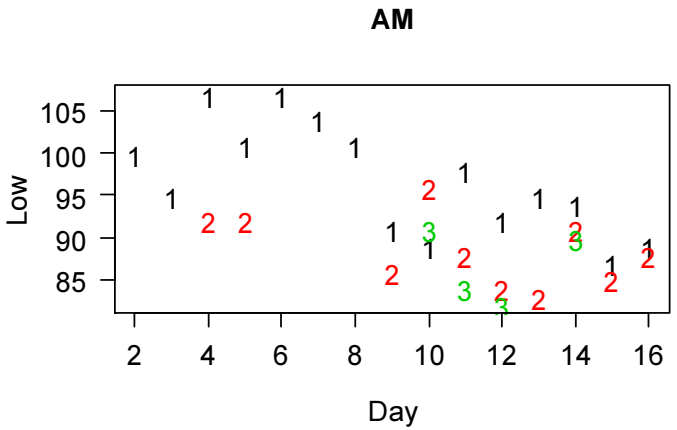
モデルの適合度

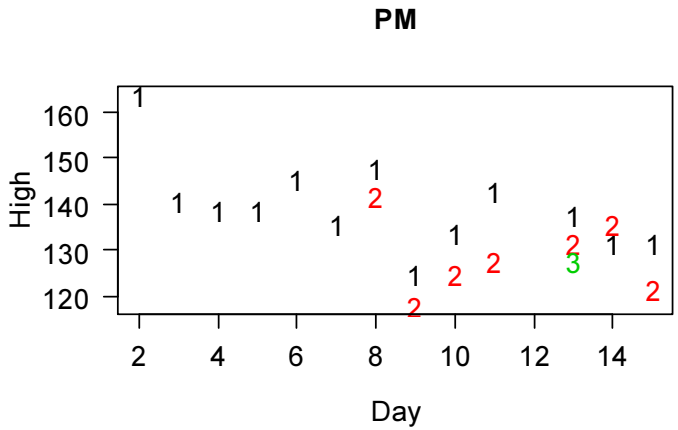

Residual
20
10
0
-10
-20
130
135
140
145
Predicted

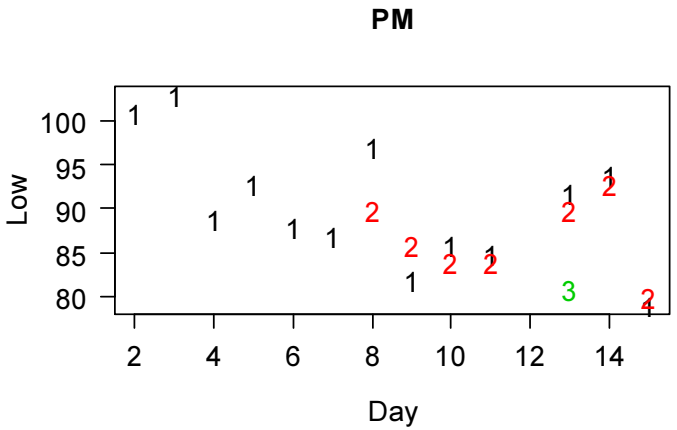

Normal Q-Q Plot
Residual
20
10
0
-10
-20
-2
-1
0
1
2
Theoretical Quantiles

重回帰モデル









重回帰例

```
modelH.3 <- lm(High~Day+AP+Iteration,data=bp.dat)

summary(modelH.3)$coef
```

重回帰注意

・線形モデルというのはパラメータに関して線形ということなので，説明変数に非線形なものが入っていてもOK

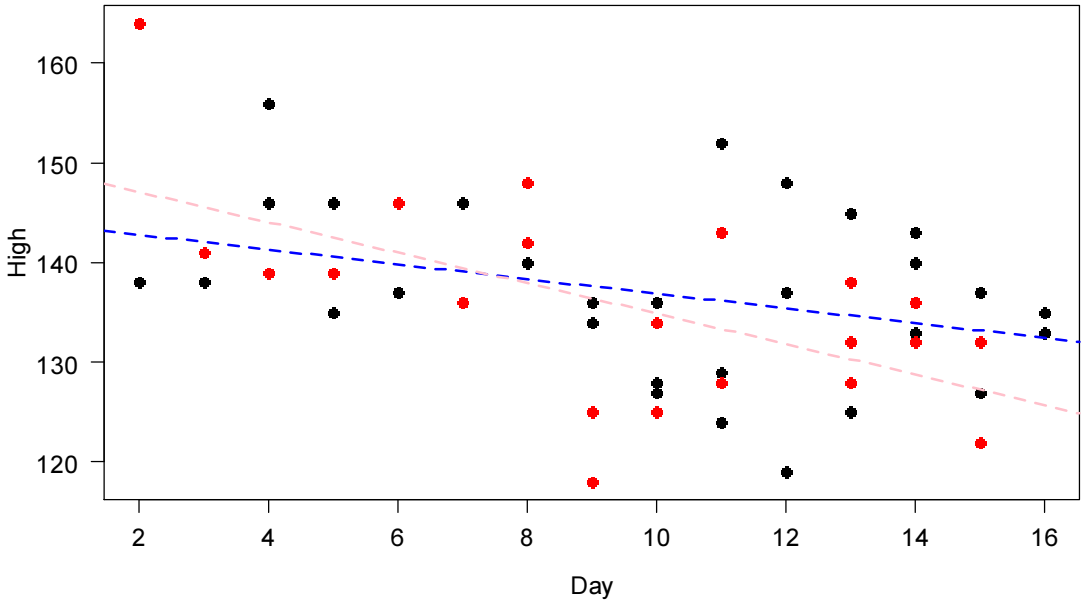
例：modelH.Poly <- lm(High~Day+I(Day^2)+I(Day^3),data=bp.dat)

summary(modelH.Poly)$coef

交互作用

・日による減少傾向は午前と午後で異なるか？

```
modelH.IA <- lm(High~Day*AP, data=bp.dat)

summary(modelH.IA)$coef
```



モデル選択 ～ AIC

$E[\log(g(x|\theta))] \approx (1/n)\, \Sigma \log(g(x|\theta))$

は悪い近似．より良い近似は，

$E[\log(g(x|\theta))] \approx (1/n)\, (\Sigma \log(g(x|\theta)) - K)$

となる（$K = \dim(\theta)$：パラメータ数）．

良いモデルはKL距離 $= E[\log(f(x))] - E[\log(g(x|\theta))]$ を最小にするものであるから，$\Sigma \log(g(x|\theta)) - K$を最大にすれば良い．

AIC = -2×対数尤度 + 2×パラメータ数

と定義するとAICを最小にするモデルが良いモデル．

AIC使用例

```
AIC(modelH.1,modelH.2,modelH.3)

modelH.f <- update(modelH.3, ~.^2)

library(MASS)
stepAIC(modelH.f)
```

AICc

・AICは大標本を仮定している
・小標本のときに使えるAIC

$$AICc = AIC + 2K(K+1)/(n - K - 1)$$

n/K < 40なら，AICcを使うべき（Burnham ＆ Anderson 2002）
正規線形モデル仮定を利用して導出しているので他のモデルではパフォーマンスが悪いかも…

AICc使用例

```
library(MuMIn)

dredge(modelH.f)

model.avg(dredge(modelH.f),subset = weight > 0.05)
```

予測

predict(modelH.b, newdata=list(Day=30,Iteration=factor(1)))

問題：

1. 何日目に正常血圧（125）に達するか？

2. 何日目に95%の確率で正常血圧より低くなるか？

デルタ法

30日後のHighとLowの比はどうなるか？

$\mathrm{var}(f(x)) = (df/dx)^2 \, \mathrm{var}(x)$

$\mathrm{var}(\mathrm{High}/\mathrm{Low}) = (1/\mathrm{Low})^2 \, \mathrm{var}(\mathrm{High}) + (-\mathrm{High}/\mathrm{Low}^2)^2 \, \mathrm{var}(\mathrm{Low})$

$= (\mathrm{High}/\mathrm{Low})^2 \, \mathrm{CV}(\mathrm{High})^2 + (\mathrm{High}/\mathrm{Low})^2 \, \mathrm{CV}(\mathrm{Low})^2$

$= (\mathrm{High}/\mathrm{Low})^2 \, \{\mathrm{CV}(\mathrm{High})^2 + \mathrm{CV}(\mathrm{Low})^2\}$

その他

- 診断

  plot(modelH.1)

  influence.measures(modelH.1)
- 切片なしモデル

  lm(High~Day-1, data=bp.dat)
- 分散分析

  anova(modelH.f)
- offset

  lm(High~offset(Low)+Day,data=bp.dat)

一般化線形モデル（GLM）

- 誤差分布が正規分布でなくても良い（二項分布，ポアソン分布，…）

例： glm(cbind(x, n-x) ~ z, family=binomial)

glm(y ~ x, family=poisson)

- 説明変数としてカテゴリカル変数も扱える

例： glm(y ~ factor(a), family=poisson)

よく使われる確率分布とリンク関数

	分布	**デフォルトのリンク関数**
離散変数	二項分布 (0/1) binomial	logit
	ポアソン分布 (0, 1, 2..) poisson	log
連続変数	正規分布 gaussian	identity
	ガンマ分布 Gamma	inverse

```
>  ? family
```

二値データ

・0/1データ　0, 1, 1, 0, 1, …（出生/死亡，釣獲/脱落，…）
・n回中x回起こった（船の出漁数，…）

```
z <- rnorm(20)
x <- rbinom(20,1,1/(1+exp(-(0.3-0.2*z))))
glm(x~z,family=binomial)
x <- rbinom(20,5,1/(1+exp(-(0.3-0.2*z))))
glm(cbind(x,5-x)~z,family=binomial)
```

カウントデータ

- 0, 1, 2, 3, …
- 魚の尾数，オットセイの群れの数，…

脈拍は実際は連続値であるが，ここでは離散データとして扱う

modelP.f <- glm(Pulse~Day+AP+Iteration,family=poisson,data=bp.dat)

過分散

- Var(X) > E(X)

負の二項分布

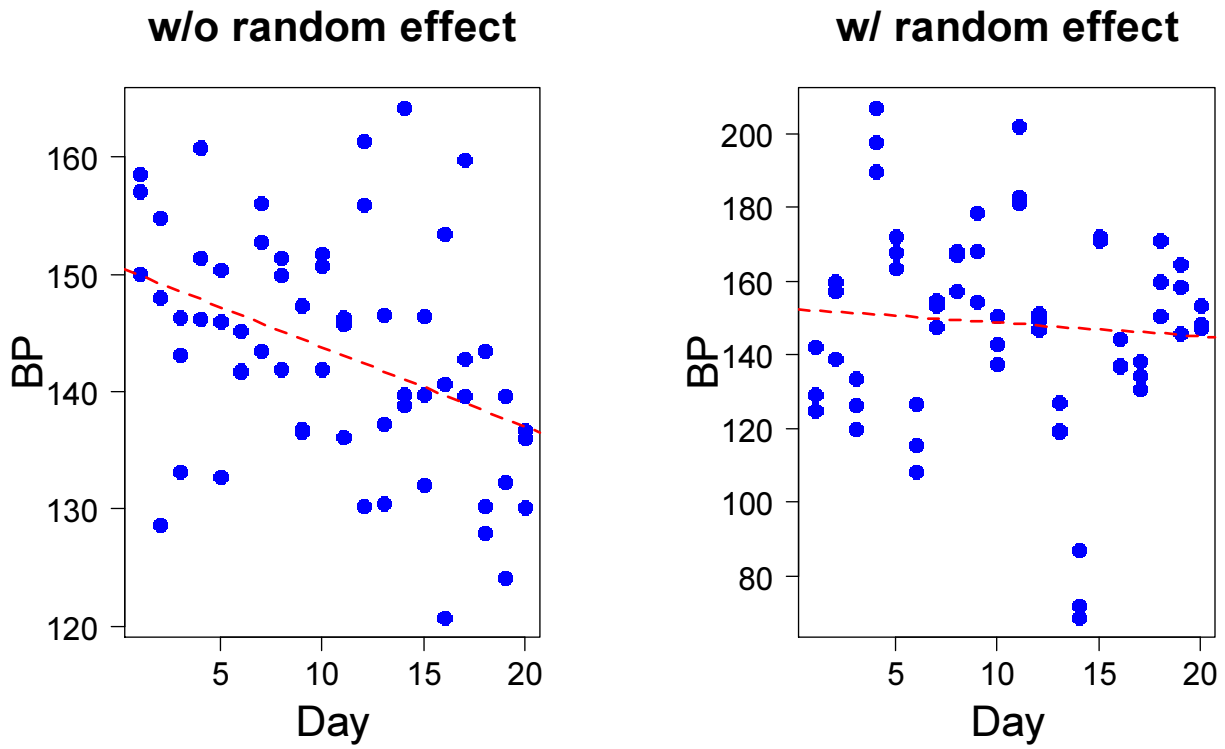
$$\Pr(x) = \frac{\Gamma(d+x)}{\Gamma(d)x!}\left(\frac{\mu}{\mu+d}\right)^{x}\left(\frac{d}{\mu+d}\right)^{d}$$

E(X)=μ, Var(X) = μ + μ^2/d

```
library(MASS)
z <- rnorm(30)
x <- rnbinom(30,size=0.5,mu=exp(0.2-0.3*z))
glm.nb(x~z)
```

ランダム効果モデル

同じ日の同じ時間帯の測定結果は同じ値を持つ傾向がある？



ランダム効果モデル

・同じ日の血圧は同じような値

$Y_{ij} = \mu_i + b \times day_{ij} + e_{ij}$, (i = 日, j = 繰り返し)

$e_{ij} \sim N(0, \sigma^2)$

$\mu_i = \mu + r_i$

$r_i \sim N(0, \sigma_r^2)$

パラメータ推定：尤度関数 $\int p(y|r)p(r)dr$ を最大化

```
library(lme4)
lmer(High ~ Day+Iterarion+(1|ID), data=bp.dat,REML=FALSE)
```

ランダム効果モデルの利点と欠点

- Type I error過小推定の回避
- 過分散を扱う
- 柔軟なモデリングを可能にする
- 潜在要因・構造を考慮
- 欠測値を扱える

- 計算が大変

GLMM

- Generalized Linear Mixed Models
- 応答変数の確率分布として正規分布以外の確率分布も扱う
- ランダム効果は通常，正規分布を仮定する

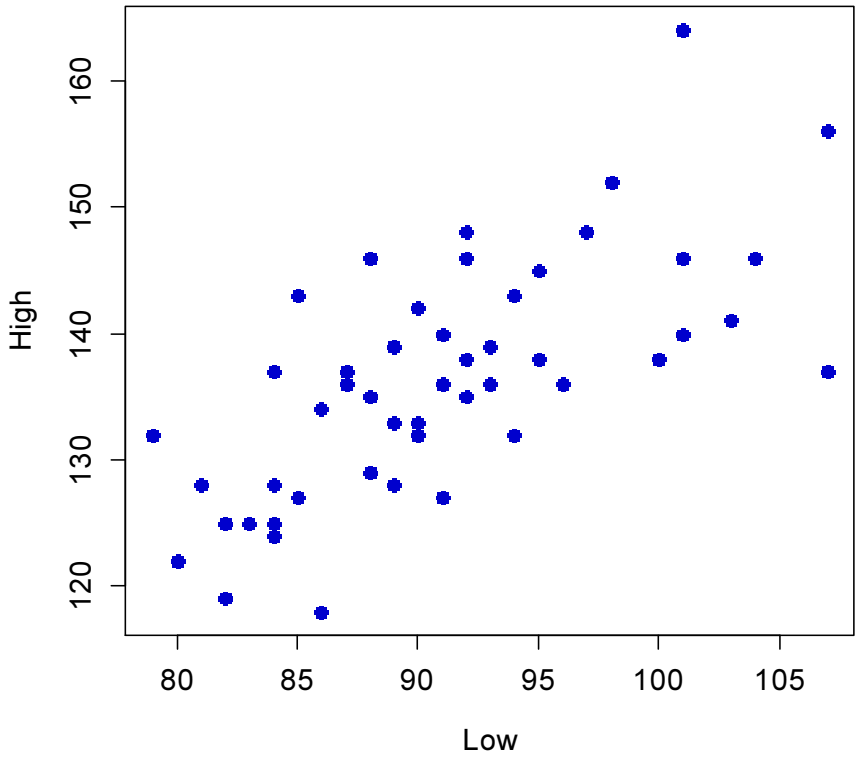
- lme4にはglmerという関数がある

  glmer((High - Low > 40) ~ Day+(1|ID),family=binomial,data=bp.dat)

- library(glmmML)なども

ベクトル回帰

- 血圧の上と下には相関がある
- 血圧の上と下の減少率は同じか，違うか？
- 血圧の上と下に朝夜の影響の違いはあるか？



ベクトル回帰

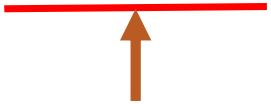
$$\begin{pmatrix} H_i \\ L_i \end{pmatrix} = \begin{pmatrix} a_H + b_H Day_i \\ a_L + b_L Day_i \end{pmatrix} + \begin{pmatrix} \epsilon_i \\ \nu_i \end{pmatrix}$$

$$\begin{pmatrix} \epsilon_i \\ \nu_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_H^2 & \rho\sigma_H\sigma_L \\ \rho\sigma_H\sigma_L & \sigma_L^2 \end{pmatrix} \right)$$

水温と漁獲量（or CPUE）の間の関係は？
複数種の関係は？

ベクトル回帰

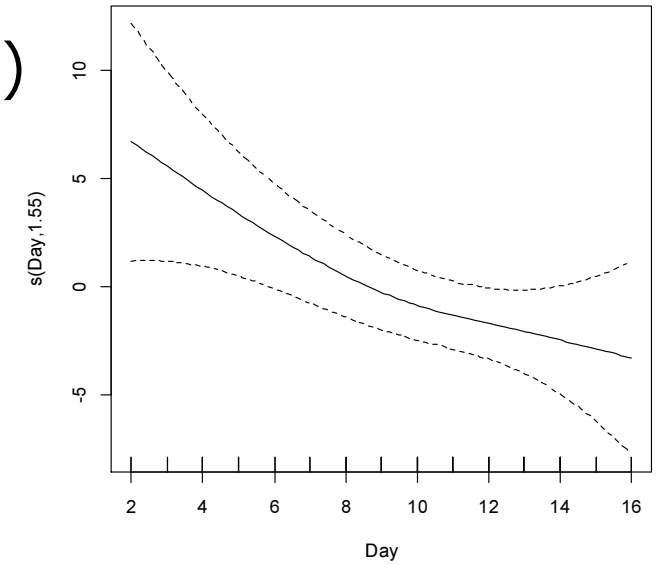
```
library(VGAM)
modelV1.1 <- vglm(cbind(High, Low)~Day+AP+Iteration, data=bp.dat,
binormal(eq.mean=FALSE), maxit=1000)

modelV1.1.4 <- vglm(cbind(High, Low)~Day+AP+Iteration, data=bp.dat,
binormal(eq.mean=~Day+AP-1), maxit=1000)
```



Dayの係数と午前/午後（AP）の効果はHighとLowで共通
切片とIterationはHighとLowで違うパラメータとして推定される

一般化加法モデル（GAM）

- ノンパラメトリック回帰
- 非線形な変化を扱える（水温，空間分布，…）



- library(mgcv)
- modelH.GAM <- gam(High~s(Day)+AP+Iteration, data=bp.dat)

GLMの応用

- 状態空間モデル（State-Space Model）

  $X_t = F(X_{t-1}, e_t)$

  $Y_t = G(X_t, v_t)$

- GLM-Tree

  Ichinokawa, M., and Brodziak, J. 2010. Fish Res 106(3): 249-260.

  http://cse.fra.affrc.go.jp/ichimomo/Tuna/glm.tree.html

- Zero-inflated Models ~ ZINBNB

  Okamura, H. et al. 2012. Population Ecology 54(3): 467-474.

ベイズ推定

- 事後確率 ～ 尤度×事前分布
- MCMC

https://sites.google.com/site/hiroshiokamura/bayes

水産資源学で使われる回帰

- CPUE標準化（線形回帰/GLM/GLMM）
- DeLury法（線形回帰）
- 死亡係数推定（線形回帰）
- 成長曲線推定（線形・非線形回帰）
- 成熟曲線推定（ロジスティック回帰）
- 個体群モデル（線形・非線形回帰/GLM/GLMM）
- 空間分布モデル（GLM/GLMM/GAM/GAMM）
- 年齢組成・体長組成（VGAM/VGAMM）
- 種間関係モデル（VGAM/VGAMM）

Homework

血圧測定しよう！

自分のデータにモデルを適用してみよう！