RとGLM入門

2007年10月16日（火），中央水研

遠洋水研外洋資源部鯨類管理研究室

岡村　寛

Rとは？

- Rは統計分析をするのに特に優れたソフトウェアである．フリーソフトなので，購入費等はかからない．
- グラフィック，プログラミングにも優れている．
- 次々と便利な最新の統計パッケージが作られ，利用可能となっている
- 他のプログラム（WinBUGS，ADMB，…）とのリンクも容易

準備

- まずはじめにDocumentフォルダなどに作業用フォルダを作ります．そこに分析したいデータを入れておきます．
- この作業用フォルダは，研究にひとつ作ります．たとえば，マイワシの研究をしていたらマイワシにR作業フォルダをひとつ，次にカタクチイワシの研究をするときにはカタクチイワシに新たに作業フォルダを作ります．
- 最初の作業が終わったら，workspaceイメージを保存．次回からは.RDataをクリックして作業開始．

注意点

- こまめにセーブ．Rは比較的安定しているが，時にクラッシュする．
- help機能が充実しているので，分らないことがあったらhelpを見る．インターネットの情報なども役に立つ．

```
> ?glm
```

データの作成

- データ作成の際の注意

  日本語は使わない

  ヘッダにはなるべく簡単な略式記号が良い

  ヘッダに*，?，/，_，. などの記号は使わない方が良い

  スペースもむやみに空けない

  変数名 〇 b3，× 3b

  扱いやすいデータにしといて，別にメモを作っておくのが後々のためにも良いだろう

変数

- Rではデータや分析結果などをひとつの変数（正確にはオブジェクト）としていろいろ操作する

```
> x <- c(1,2,3)
> x
[1] 1 2 3
> x <- matrix(c(1,2,3,4),ncol=2)
> x
     [,1] [,2]
[1,]    1    3
[2,]    2    4
```

計算

```
> x <- c(1,2,3,4)
> y <- 8:5
> x+y
[1] 9 9 9 9

> x <- matrix(rnorm(4*4),ncol=4)
> x%*%y
          [,1]
[1,]  15.23455
[2,]  11.45978
[3,] -24.07344
[4,]  10.52256
```

要約統計量

```
> x <- c(1,2,3,4)
> mean(x)
[1] 2.5
> var(x)
[1] 1.666667
> sd(x)
[1] 1.290994
> median(x)
[1] 2.5
> c(min(x),max(x))
[1] 1 4
> quantile(x,prob=0.2)
20%
1.6
```

データのインポート

- csvデータを読み込む場合
  data1 <- read.csv(“data1.csv”)
- txtデータを読み込む場合
  data1 <- read.table(“data1.txt”,header=T)
- 桁数を指定して読み込む
  data1 <- read.fwf(“data1.dat”,width=c(3,-1,2))

Excelファイルを直接読み込むことも可能だがそれほど必要ないだろう. scanなども便利.

データのエクスポート

- csvでデータを保存する場合

  write.csv(res1,“res1.csv”)

- txtでデータを保存する場合

  write.table(res1,“res1.txt”)

cat(res1, file = “res.dat”, sep =“ ”)

データフレーム

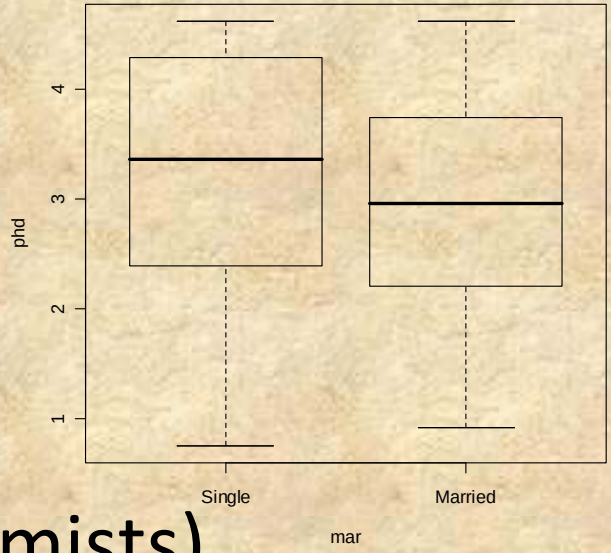
```
> bioChemists
   art fem     mar     kid5  phd  ment
1  0   Men     Married 0     2.52 7
2  0   Women   Single  0     2.05 6
3  0   Women   Single  0     3.75 6
4  0   Men     Married 1     1.18 3
5  0   Women   Single  0     3.75 26
```

データフレームの使い方を覚えるといろいろ便利

```
> lm(phd ~ mar, data=bioChemists)

Coefficients:

(Intercept)   marMarried
     3.2165      -0.1712
```




```
> plot(phd ~ mar, data=bioChemists)
```

データの種類と想定される統計モデル

- 連続データ

  -∞～∞　正規分布，t分布

  0～∞（0含まず）　対数正規分布，ガンマ分布

  0～1　ベータ分布

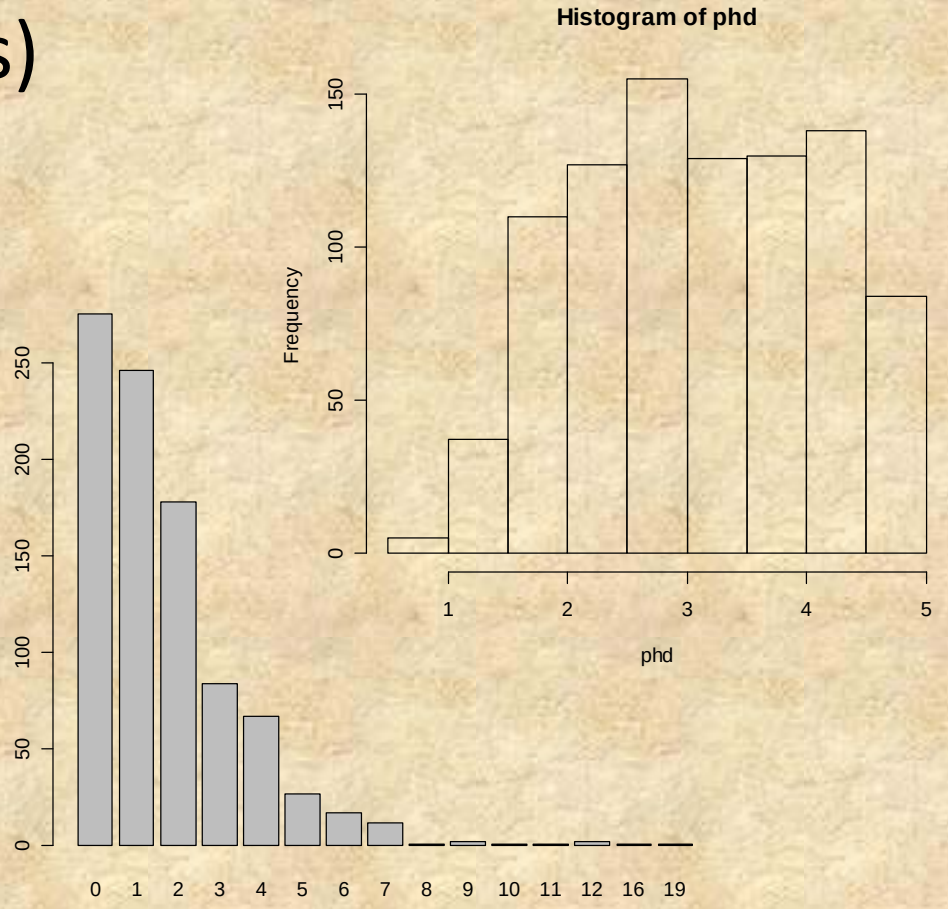
- 離散データ（カウントデータ）

  0 or 1　二項分布

  0, 1, …, ∞　ポアソン分布，負の二項分布

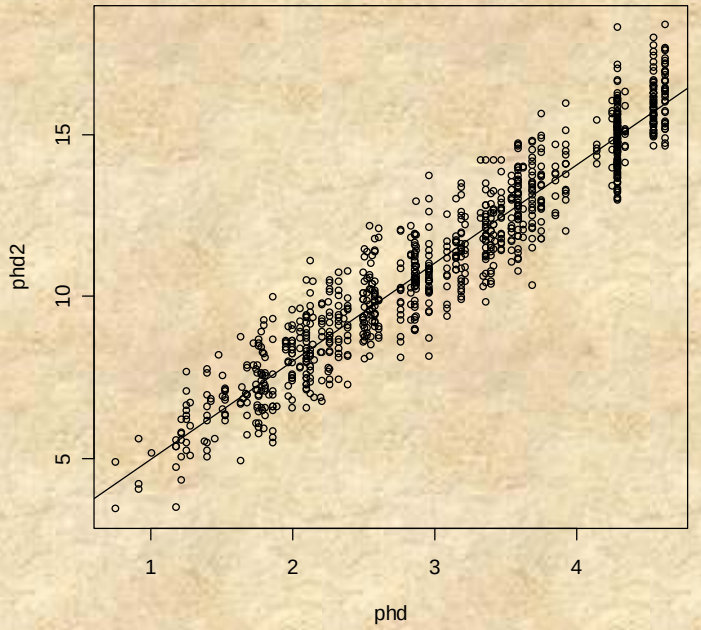
データの図示

```
> attach(bioChemists)
> hist(phd)
> barplot(table(art))
```



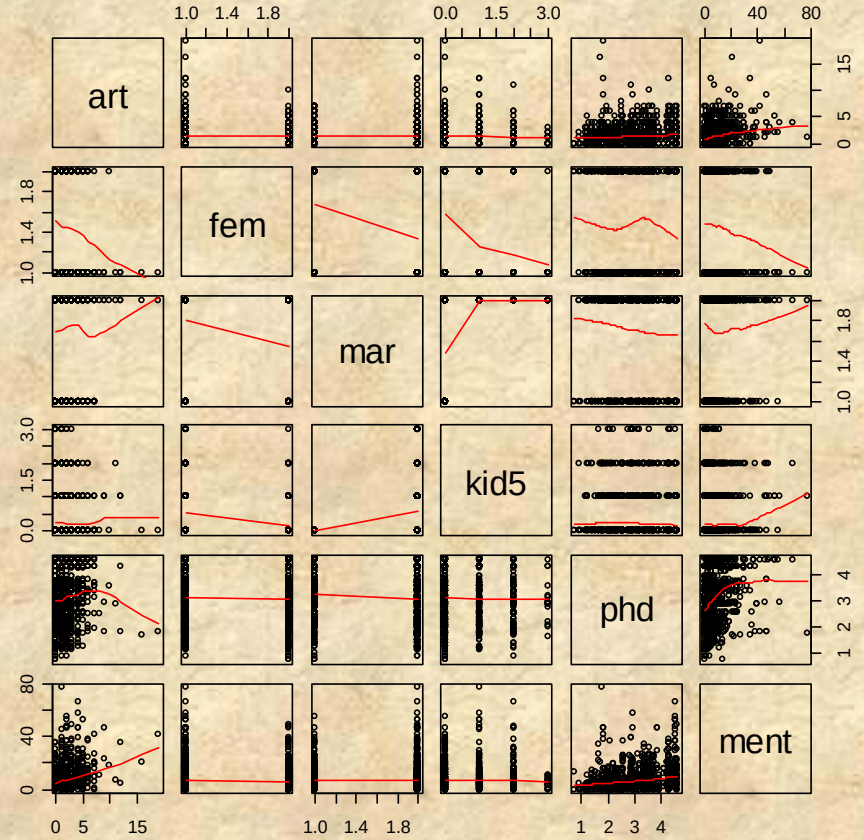
データの図示

```
> phd2 <- 3*phd + 2 +rnorm(length(phd))
> plot(phd,phd2)
> fm1 <- lm(phd2~phd)
> abline(fm1)
```



データの図示

- pairs(bioChemists,panel=panel.smooth)



統計分布

```
> dnorm(0,0,0.1)          # 確率密度
[1] 3.989423
> plnorm(2,log(2),0.3)          # 分布関数
[1] 0.5
> qt(0.95,6)                    # 分位点
[1] 1.943180
> rpois(5,2)                     # 乱数生成
[1] 5 0 2 1 4
```

統計分析

```
> x1 <- c(1,2,3,4)
> x2 <- c(3,2,5,6)
> t.test(x1,x2)
t = -1.3416, df = 5.4, p-value = 0.2334

> x <- matrix(c(6,15,9,2),ncol=2)
> chisq.test(x)

X-squared = 6.2196, df = 1, p-value = 0.01263
```

直線回帰

ある値y（たとえば，漁獲量）は水温によってどう変わるか？

```
> x <- rnorm(10,5,1)
> y <- rlnorm(10,log(10+0.5*x),0.3)
> summary(lm(y~x))

Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  0.04409    5.32656      0.008     0.9936
x            2.34539    0.92612      2.533     0.0351 *
```

直線回帰の応用

個体数の時系列

> x

[1] 117 105 130 122 136 129 129 125 131 163

がある．プロダクションモデルをフィットして，増加率r，環境収容量Kを推定せよ．

解答

$N_{t+1} = N_t + rN_t(1-N_t/K)$なので，右辺を整理すると，

$N_{t+1} = (r+1)N_t-(r/K)N_t^2$となる．

```
> lm(x[2:10]~x[1:9]+I(x[1:9]^2)-1)
Coefficients:
     x[1:9]  I(x[1:9]^2)
   1.658478    -0.004922
```

r = 0.658,　K = 0.658/0.004922 = 133.8

非線形モデル

```
> f1 <- function(p,x) sum((log(x[2:10])-log(p[1]*x[1:9]+p[2]*x[1:9]^2))^2)


> optim(c(1,0),f1,x=x)
$par
[1]  1.677515694 -0.005116905

$value
[1] 0.09825909
```

直線回帰の問題

- 応答変数(y)が漁獲量であるとすると正の値だけをとるべきである．正規誤差を仮定すると信頼区間や予測値は負の値もとりうる．
- 応答変数(y)が漁獲尾数であるとすると正の値であり，しかも整数である．正規分布では余計都合が悪い．
- 説明変数(x)がカテゴリーである場合（分散分析）も扱いたい．

一般化線形モデル

- 直線回帰，分散分析の一般化
- 正規分布以外の分布（二項分布，ポアソン分布）などが扱える（これらを指数分布族という）
- 説明変数は連続であっても，カテゴリーであってもどちらもOK
- 0-1データ →　二項分布

  カウントデータ（ランダム）　→　ポアソン分布

  カウントデータ（集中）　→　負の二項分布

リンク関数

- 正規分布

  $\mu = a_0 + a_1 x_1 + a_2 x_2 + \ldots$

- 二項分布

  $\mathrm{logit}(p) = \log\{p/(1-p)\} = a_0 + a_1 x_1 + a_2 x_2 + \ldots$

- ポアソン分布，負の二項分布

  $\log(\lambda) = a_0 + a_1 x_1 + a_2 x_2 + \ldots$

たとえば二項分布の場合

```
> glm(as.numeric(mar)-1~fem,data=bioChemists)
Coefficients:
(Intercept)    femWomen
    0.7713      -0.2368                          AIC: 1173

> glm(as.numeric(mar)-1~fem,family=binomial,data=bioChemists)
Coefficients:
(Intercept)    femWomen
     1.215      -1.077                           AIC: 1117
```

女と男の結婚率は？

```
> 1/(1+exp(-(1.215-1.077)))    # Women
[1] 0.5344454
> 1/(1+exp(-(1.215)))          # Men
[1] 0.7711824
```

ポアソン回帰

```
> glm(art~mar,family=poisson)

Coefficients:
(Intercept)   marMarried
    0.46514      0.09117
AIC: 3486
```

負の二項分布回帰

```
> library(MASS)
> glm.nb(art~mar)

Coefficients:
(Intercept)   marMarried
    0.46514      0.09117
AIC: 3224
```

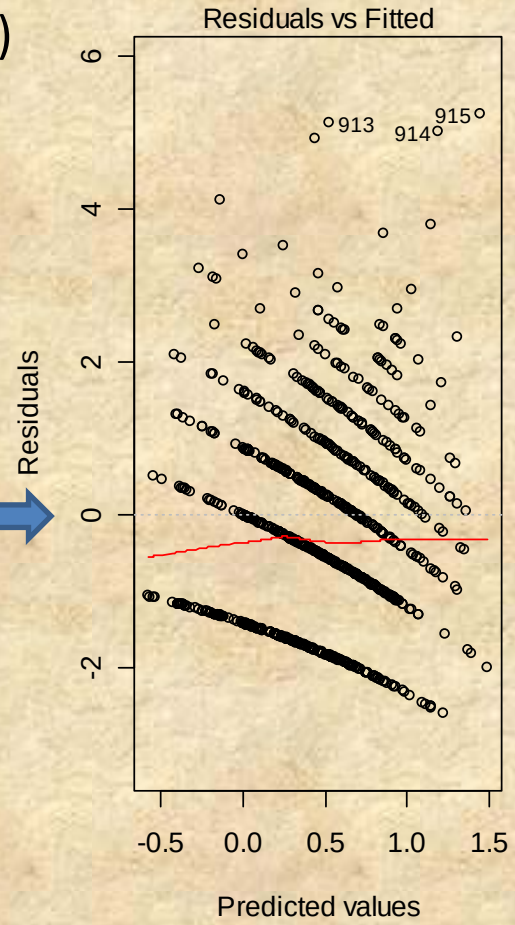
詳しい結果を見る

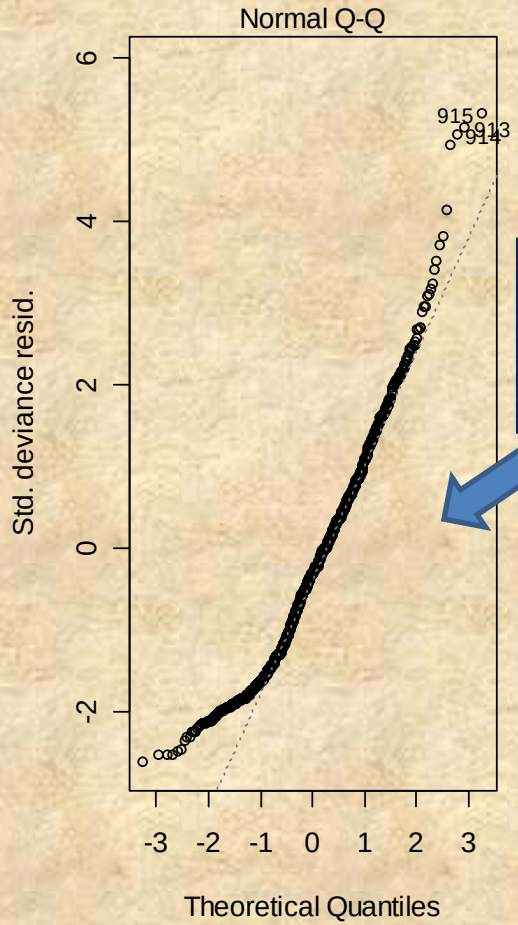
```
> res <- glm.nb(art~mar)
> summary(res)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.46514    0.06263   7.427 1.11e-13 ***
marMarried   0.09117    0.07636   1.194    0.233
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

モデル診断

```
> res <- glm(art~fem+mar+kid5+phd+log(ment+1),family=poisson)
> plot(res,which=1:2)
```

0のまわりで大体均等にばらついていると良い





直線上にのってると良い

offset

- たとえば漁獲尾数が何で変わるかを知りたいとき，あらかじめ努力量の大きさの違いを考慮に入れておきたいといったようなことがある．この場合，offsetを使う．

```
> glm(catch ~ offset(log(Effort))+SST+…,
  family=poisson)
```

Cを見つつも，C/E ~ SST+…を見ていることになる

ランダム効果

- $y_{ij} = mu + a_i + e_{ij}$

a_iの効果を入れたいが，それぞれの効果には興味がないことがある．たとえば，y_{ij}は魚の体長，a_iはサンプル場所とする．サンプル場所間で差があるかどうかには興味があるが，個々のサンプル場所でどこが大きいか小さいかに興味があるわけではない場合がある．たとえば，サンプル場所が100ヶ所で，場所ごとに体長を測った魚が3尾だったすると，パラメータ数は100個になって，個々のa_iの推定値にはあまり信頼性がおけないであろう．そういうとき，個々のa_iに興味があるわけではないのに…と困る．そんなときランダム効果の考えを使うとうまくいく．

ランダム効果を使って役立つ場合

- パラメータが非常に多くなるとき，それらをランダム効果とみなせるなら，パラメータ節約的なモデルを作れる
- パラメータが変動していると考えた方が自然なとき，隠れ変数のようなものを扱いたいとき
- 個体差等が大きいときにその変動を取り込む
- missing dataがある場合
- データ間の相関をうまく取扱いたいとき

LMM

```
> library(lme4)
> glm(phd~factor(art))
Coefficients:
 (Intercept)   factor(art)1   factor(art)2   factor(art)3   factor(art)4
     2.98473       0.04255       0.25303       0.20444       0.17714
 factor(art)5   factor(art)6   factor(art)7   factor(art)8   factor(art)9
     0.39787       0.51527       0.63361      -0.47473      -0.57473
factor(art)10  factor(art)11  factor(art)12  factor(art)16  factor(art)19
     0.60527      -0.12473       0.09027      -1.24473      -1.12473
AIC: 2575
```

LMM

```
> lmer(phd~1|factor(art),method="ML")
Linear mixed-effects model fit by maximum likelihood
Formula: phd ~ 1 | factor(art)
  AIC  BIC logLik MLdeviance REMLdeviance
 2568 2577  -1282       2564        2568
Random effects:
 Groups      Name        Variance  Std.Dev.
 factor(art) (Intercept) 0.0092178 0.09601
 Residual                0.9590305 0.97930
number of obs: 915, groups: factor(art), 15

Fixed effects:
            Estimate Std. Error t value
(Intercept)  3.14159    0.05182   60.63
```

GLMM

```
> lmer(art ~ 1|factor(ment), data = dat, family="poisson",method="Laplace")
Generalized linear mixed model fit using Laplace
Formula: art ~ 1 | factor(ment)
   Data: dat
 Family: poisson(log link)
  AIC  BIC logLik deviance
 1674 1683 -834.9     1670
Random effects:
 Groups       Name        Variance Std.Dev.
 factor(ment) (Intercept) 0.19077  0.43677
number of obs: 915, groups: factor(ment), 49

Estimated scale (compare to  1 )  1.317195

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.76278    0.07534   10.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GAM

- 線形関係で必ずしも表現できない場合がある．しかし，カテゴリー変数で扱うとパラメータがやたら増えるし，カテゴリー分けが主観的になってしまうという問題がある．そういうとき，ノンパラメトリック回帰を使うとうまくいくことがある．
- GAMを使って役立つ時：

  時空間モデル

  水温効果等
- Rでは，mgcvが標準か．VGAM（応答変数が多変量の場合にも使える）なども良いだろう．gamlssも．

GAM

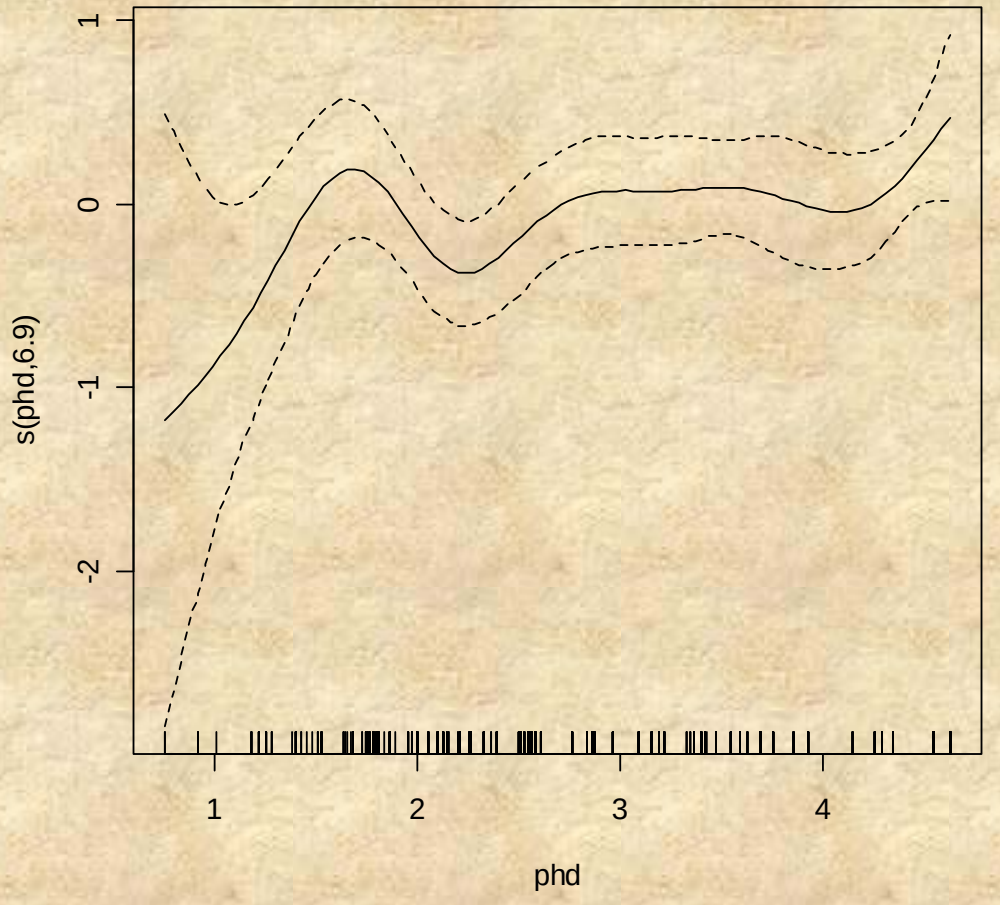
> library(mgcv)
This is mgcv 1.3-25
> gam(art~s(phd))

Family: gaussian
Link function: identity

Formula:
art ~ s(phd)

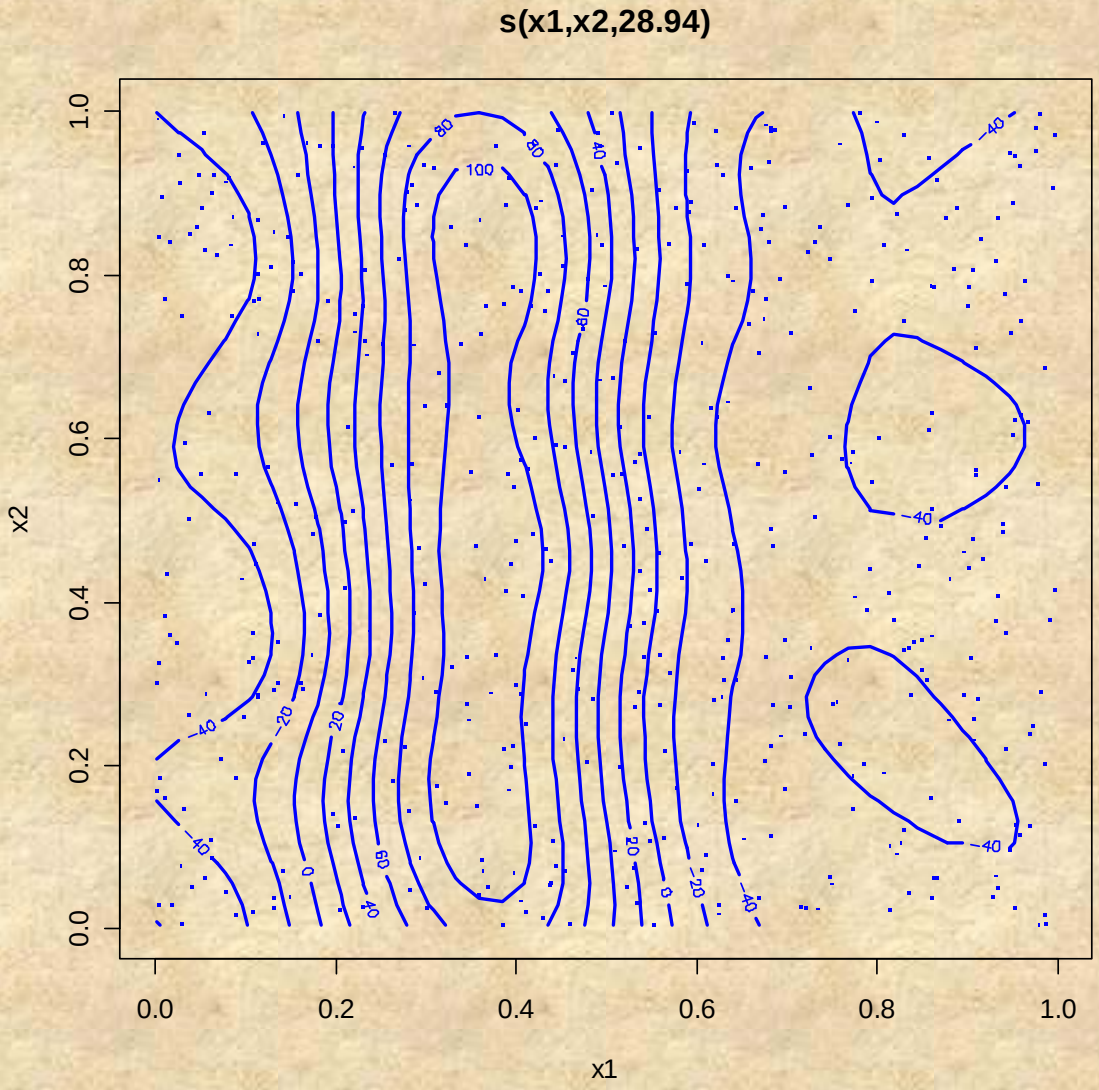
Estimated degrees of freedom:
6.897915 total = 7.897915

GCV score: 3.700284
> plot(gam(art~s(phd)))



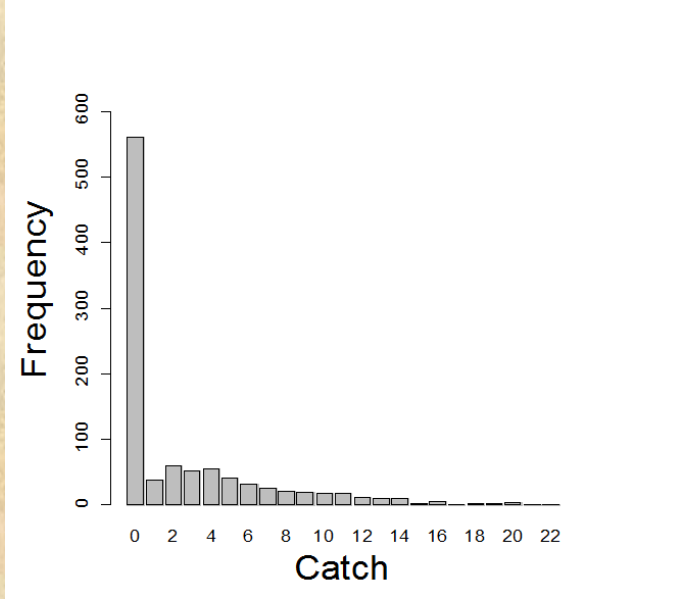
GAM

```
x1 <- runif(500,0,1)
x2 <- runif(500,0,1)
f1 <- function(x,y)
    10*(10*x)^6*(1-
    x)^11 + exp(0.5*y)
z <- f1(x1,x2)
b1 <- gam(z~s(x1,x2))
plot(b1,se=F,col=2,lwd
    =2,cex=2)
```



Zero-inflated model

- カウントデータでover-dispersionは負の二項分布でモデル化することが多いが，しばしばそれ以上にover-dispersionしている場合がある．特に0データが極端に多いとき（混獲データとか）．



Zero-inflated model

zero-inflated model (Lambert 1992, Hall 2000, Martin et al. 2005, Minami et al. 2007 など)

$$f(y) = \begin{cases} p + (1-p)q(0|m,\theta) & y = 0 \\ (1-p)q(y|m,\theta) & y = 1, 2, \ldots \end{cases}$$

平均 (1-p)m

Zero-inflated model

```
> library(pscl)
> zeroinfl(art ~ .|., data = bioChemists)

Call:
zeroinfl(formula = art ~ . | ., data = bioChemists)


Count model coefficients (poisson with log link):
(Intercept)    femWomen   marMarried        kid5         phd        ment
    0.64087    -0.20917     0.10374    -0.14328    -0.00612     0.01809

Zero-inflation model coefficients (binomial with logit link):
(Intercept)    femWomen   marMarried        kid5         phd        ment
 -0.577121    0.109730   -0.354087    0.217284    0.001412   -0.134175
```

Zero-inflated model

```
> zeroinfl(art ~ .|., data = bioChemists,  dist = "negbin",EM=TRUE)

Call:
zeroinfl(formula = art ~ . | ., data = bioChemists, dist = "negbin",
    EM = TRUE)


Count model coefficients (negbin with log link):
(Intercept)    femWomen   marMarried        kid5         phd        ment
  0.4167583   -0.1954955    0.0975894   -0.1517240   -0.0006968    0.0247846
Theta = 2.6549

Zero-inflation model coefficients (binomial with logit link):
(Intercept)    femWomen   marMarried        kid5         phd        ment
   -0.19246     0.63599     -1.49870     0.62832     -0.03755     -0.88191
```

Hurdle model

hurdle model (Welsh et al. 1996, O'Neill and Faddy 2003, Cunningham and Lindenmayer 2005)

$$f(y) = \begin{cases} p & y = 0 \\ (1-p)q(y|m,\theta)/(1-q(0|m,\theta)) & y = 1, 2, \ldots \end{cases}$$

平均 $(1-p)m/(1-q(0|m,\theta))$

zero部分とnon-zero部分を別々に計算できる

Hurdle model

```
> hurdle(art ~ ., data = bioChemists,  dist = "negbin")

Call:
hurdle(formula = art ~ ., data = bioChemists, dist = "negbin")


Count model coefficients (truncated negbin with log link):
(Intercept)    femWomen   marMarried        kid5         phd        ment
   0.355110   -0.244666    0.103416   -0.153262   -0.002940    0.023740
Theta = 1.8284

Zero hurdle model coefficients (binomial with logit link):
(Intercept)    femWomen   marMarried        kid5         phd        ment
    0.23680    -0.25115     0.32623    -0.28525     0.02222     0.08012
```

より複雑なモデル

より複雑なモデル　→　最尤法，階層モデル

最尤法：　自分でプログラムを書いてoptimなどで最適化．パラメータが多い，計算時間がかかるときなどはADMB（有料）とリンクさせると良い．

> shell(“test”,invisible=T)

階層ベイズモデル：

http://cse.fra.affrc.go.jp/okamura/bayes/index.html

まとめ

- 自分のデータで興味の問題を分析してみることが理解を早める
- 不明点はまわりのR使いに聞くのが良い
- 書籍やインターネットの情報が役に立つ
- プログラムが書けると作業が効率よくなる（データを整理してアウトプットしたり，図を描いたりでも必要）

参考文献

- Ecological Modelling 157 (2-3) (2002)は，GLM，GAM（VGAMを含む）の特集号
- Fisheries Research 70 (2-3) (2004)も，GLMの特集号．こちらは漁業データ（特にCPUE標準化）を扱っているので，より分りやすいかも．特に，CPUE標準化に興味ある人には，Maunder and Puntのreviewは必読．
- Zeilieis, A. et al. 2007. Regression models for count data in R.

  http://cran.r-project.org/doc/vignettes/pscl/countreg.pdf