

割合データ, 二値データをGLMで解析する

山口 典之 (立教大・理・生命理学)

割合データ

- 性比 (雄数 / 全個体数)
- けんかの勝率 (勝ち数 / 闘争回数)
- 孵化率 (未孵化卵数 / クラッチサイズ)
- 成幼比 (成鳥数 / 全個体数)

鳥屋の皆さんがよく出会うデータ構造

従属変数が割合データのとときに生じる問題

- ・ **分母値が一定でない場合, データ点の重みが異なる.**
二回コイントスして一回表という試行 ($1/2=0.5$)
百回コイントスして五十回表という試行 ($50/100=0.5$)

どちらもデータとしては0.5だが, 重み
(0.5であるという確からしさ) が異なる.

従属変数が割合データのとときに生じる問題

- 誤差が正規分布しない (二項分布)
- 等分散ではない ($\sigma^2=npq$)
- 従属変数に上限下限がある ($0 \leq y \leq 1$)
- しばしばoverdispersion (過分散) が生じる

普通に回帰できない

**割合データの解析には
やっかいな問題がいっぱい**

ではどうするか？

昔：変数変換（アークサインなど角度変換）

ちょい前：ロジスティック回帰

いま：一般化線形モデル (logit link, binomial error)

呼び方が違うだけ



昔のことは忘れましょう

いろいろ問題あるから

GLM (logit link, binomial error)の構造

$$\ln\left(\frac{p}{1-p}\right) = a + \beta x + \varepsilon$$

link function: logit

linear predictor

error: binomial

最尤法を用いてパラメータ推定 (←各データ点の
分母の重みの違いについても考慮されている)

overdispersionがあるときには誤差分布を変えたり,
混合効果を取り入れたりして対応.

解析例 ー親の形質とブルード性比の関係ー

解析したいこと：

ヤマガラにおいて、

親の各形質とブルード性比

に関係があるか？

解析例 一親の形質とブルード性比の関係一

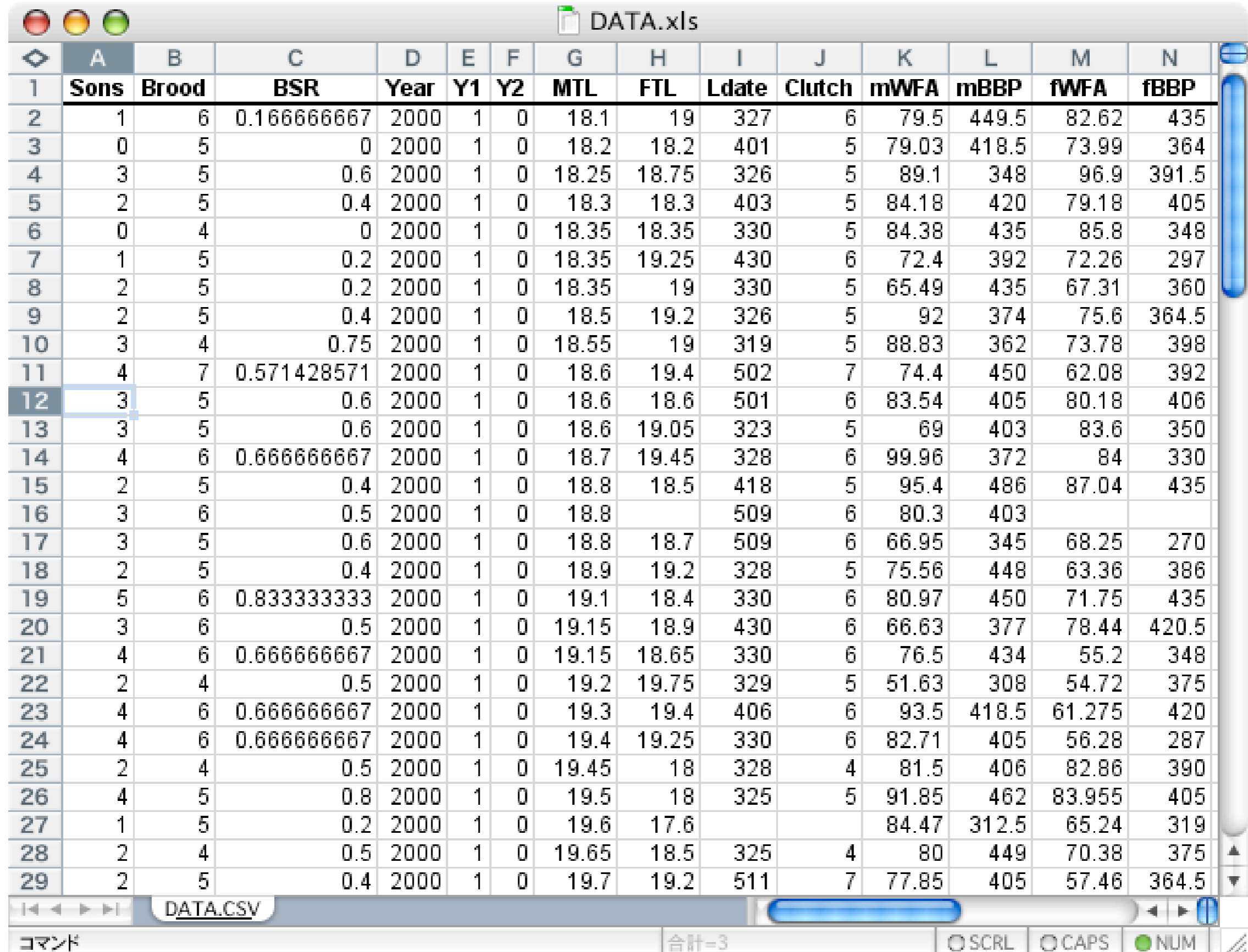
従属変数：

ブルード性比 (雄数/ブルードサイズ)

独立変数：

雄ふしよ長, 雄の胸黒斑, 雄の額白斑, 雌ふしよ長, 雌の胸黒斑, 雌の額白斑, 初卵日,
調査年 (ブロック)

データの入力 (例えばエクセルで)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Sons	Brood	BSR	Year	Y1	Y2	MTL	FTL	Ldate	Clutch	mWFA	mBBP	fWFA	fBBP
2	1	6	0.166666667	2000	1	0	18.1	19	327	6	79.5	449.5	82.62	435
3	0	5	0	2000	1	0	18.2	18.2	401	5	79.03	418.5	73.99	364
4	3	5	0.6	2000	1	0	18.25	18.75	326	5	89.1	348	96.9	391.5
5	2	5	0.4	2000	1	0	18.3	18.3	403	5	84.18	420	79.18	405
6	0	4	0	2000	1	0	18.35	18.35	330	5	84.38	435	85.8	348
7	1	5	0.2	2000	1	0	18.35	19.25	430	6	72.4	392	72.26	297
8	2	5	0.2	2000	1	0	18.35	19	330	5	65.49	435	67.31	360
9	2	5	0.4	2000	1	0	18.5	19.2	326	5	92	374	75.6	364.5
10	3	4	0.75	2000	1	0	18.55	19	319	5	88.83	362	73.78	398
11	4	7	0.571428571	2000	1	0	18.6	19.4	502	7	74.4	450	62.08	392
12	3	5	0.6	2000	1	0	18.6	18.6	501	6	83.54	405	80.18	406
13	3	5	0.6	2000	1	0	18.6	19.05	323	5	69	403	83.6	350
14	4	6	0.666666667	2000	1	0	18.7	19.45	328	6	99.96	372	84	330
15	2	5	0.4	2000	1	0	18.8	18.5	418	5	95.4	486	87.04	435
16	3	6	0.5	2000	1	0	18.8		509	6	80.3	403		
17	3	5	0.6	2000	1	0	18.8	18.7	509	6	66.95	345	68.25	270
18	2	5	0.4	2000	1	0	18.9	19.2	328	5	75.56	448	63.36	386
19	5	6	0.833333333	2000	1	0	19.1	18.4	330	6	80.97	450	71.75	435
20	3	6	0.5	2000	1	0	19.15	18.9	430	6	66.63	377	78.44	420.5
21	4	6	0.666666667	2000	1	0	19.15	18.65	330	6	76.5	434	55.2	348
22	2	4	0.5	2000	1	0	19.2	19.75	329	5	51.63	308	54.72	375
23	4	6	0.666666667	2000	1	0	19.3	19.4	406	6	93.5	418.5	61.275	420
24	4	6	0.666666667	2000	1	0	19.4	19.25	330	6	82.71	405	56.28	287
25	2	4	0.5	2000	1	0	19.45	18	328	4	81.5	406	82.86	390
26	4	5	0.8	2000	1	0	19.5	18	325	5	91.85	462	83.955	405
27	1	5	0.2	2000	1	0	19.6	17.6			84.47	312.5	65.24	319
28	2	4	0.5	2000	1	0	19.65	18.5	325	4	80	449	70.38	375
29	2	5	0.4	2000	1	0	19.7	19.2	511	7	77.85	405	57.46	364.5

コマンド 合計=3 SCRL CAPS NUM

解析例 —親の形質とブルード性比の関係—

R へのデータの読み込み

```
> para <- read.table("data.csv", header=T, sep=",")
> names(para)
[1] "Sons"      "Brood"     "BSR"       "Year"      "Y1"
"Y2"       "MTL"       "FTL"       "Ldate"
[10] "Clutch"   "mWFA"      "mBBP"      "fWFA"      "fBBP"
```

解析例 —親の形質とブルード性比の関係—

GLM解析 (Full model)

```
> model00 <- glm(cbind(Sons, Brood - Sons) ~ MTL +  
FTL + Ldate + mWFA + mBBP + fWFA + fBBP + MTL:Y1 +  
FTL:Y1 + Ldate:Y1 + mWFA:Y1 + mBBP:Y1 + fWFA:Y1 +  
fBBP:Y1 + MTL:mWFA + MTL:mBBP + FTL:fWFA + FTL:fBBP  
+ MTL:fWFA + MTL:fBBP + FTL:mWFA + FTL:mBBP, family  
= binomial, data = para)
```

誤差分布の指定

glm()関数

データ指定

従属変数

独立変数

Full modelの解析結果

```
> summary(model00)
```

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.779e+02	1.326e+02	1.341	0.1799
MTL	-9.098e+00	4.491e+00	-2.026	0.0428 *
FTL	-3.290e-01	4.528e+00	-0.073	0.9421
...
FTL:mWFA	3.066e-02	5.377e-02	0.570	0.5685
FTL:mBBP	-6.464e-03	9.047e-03	-0.714	0.4749

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter)  
Null deviance:
```

ここからoverdispersionの程度
(dispersion parameter)を計算する

```
Residual deviance: 23.526 on 21 degrees of freedom
```

```
AIC: 153.01
```

```
Number of Fisher Scoring iterations: 4
```

model selection

```
> model01 <- update(model00, ~. - Ldate:Y1)
```

```
> anova(model00, model01, test="Chi")
```

Analysis of Deviance Table

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	21	23.5257			
2	22	23.5257	-1	-0.0001	0.9939

```
> summary(model01)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.778e+02	1.322e+02	1.345	0.1787
MTL	-9.100e+00	4.485e+00	-2.029	0.0424 *
FTL	-3.238e-01	4.479e+00	-0.072	0.9424
⋮	⋮	⋮	⋮	⋮
FTL:mWFA	3.044e-02	4.572e-02	0.666	0.5056
FTL:mBBP	-6.463e-03	9.047e-03	-0.714	0.4749

Final model

```
> summary(model32)
```

Call:

```
glm(formula = cbind(Sons, Brood - Sons) ~ MTL + FTL,  
family = binomial, data = para)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
雄のふしょ長とブルード内の雄ヒナの割合に正の関係				0.00685	***
MTL	0.8243	0.2438	3.381	0.000723	***
FTL	0.5038	0.2789	1.807	0.070810	.

雌のふしょ長とブルード内の雄ヒナの割合に正の関係 of freedom

Residual deviance: 43.025 on 44 degrees of freedom

AIC: 136.51

Analysis of DevianceにもとづくP値出力

```
> anova(model32, test="Chi")
```

```
Analysis of Deviance Table
```

```
Model: binomial, link: logit
```

```
Terms added sequentially (first to last)
```

↑シーケンシャル (変数入力順を入れ替えるとP値が変わる) なので注意

	Df	Deviance	Resid.	Df	Resid. Dev	P(> Chi)
NULL				46	57.508	
MTL	1	11.168		45	46.340	0.001
FTL	1	3.315		44	43.025	0.069

```
> anova(model32.2, test="Chi")
```

```
Analysis of Deviance Table
```

	Df	Deviance	Resid.	Df	Resid. Dev	P(> Chi)
NULL				46	57.508	
FTL	1	2.301		45	55.207	0.129
MTL	1	12.182		44	43.025	0.0004825

調整済みAnalysis of Deviance

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi)
NULL			46	57.508	
MTL	1	12.182	44	43.025	0.0004825
FTL	1	3.315	44	43.025	0.069

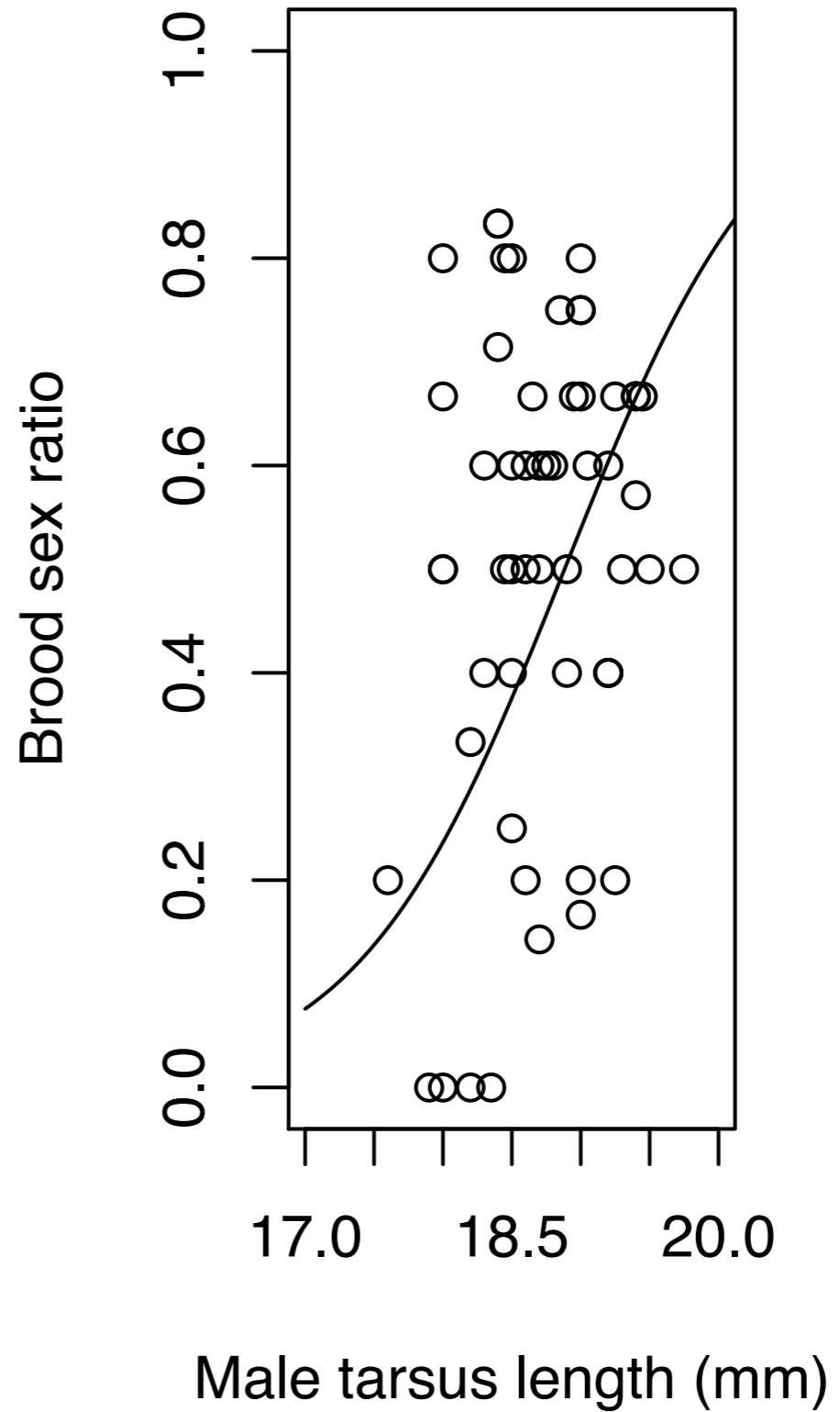
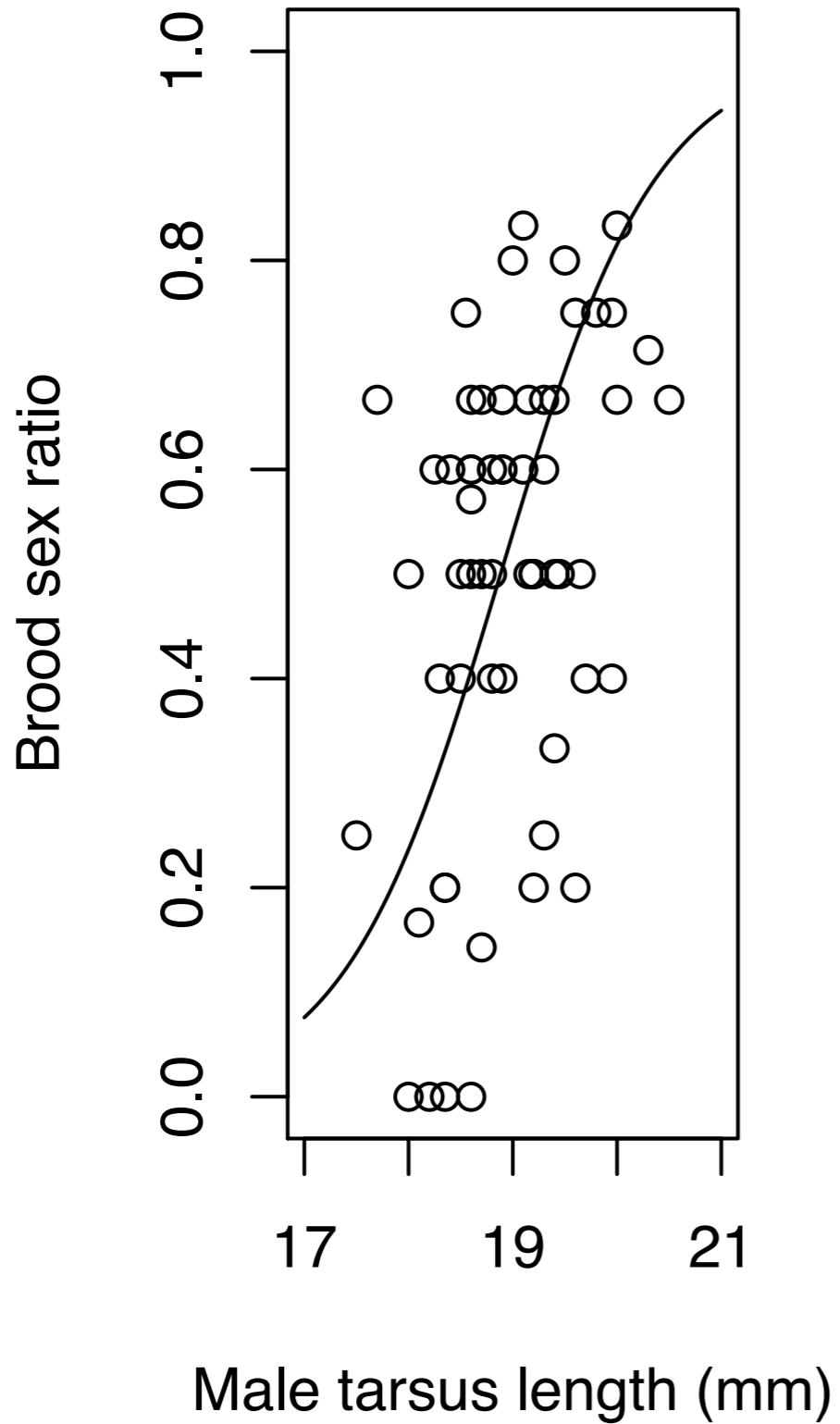
雄ふしょ長

有意

雌ふしょ長

マージナル

散布図はこんな感じ



Overdispersionへの対処法

対処しないと、どうなるか？

→ **第一種の過誤**を犯しやすくなります。

過分散の程度にもよるが、 $\alpha=0.05$ で検定しているつもりが、 $\alpha=0.1\sim 0.2$ ぐらいの甘い検定になるかも。

↑君にも出来るごまかし？

overdispersionとは…

誤差分布として**二項分布**で期待されるよりも**残差のばらつきが大きくなっている状態。**

どうしてそんな状態になるの？

—**腹卵**のようなクラスターごとに**二項分布の成功確率が異なる場合**に生じる。

→**個体差がある生き物を扱う場合にはほとんど宿命。**

ロジスティックモデルの持病のようなもの

Overdispersionへの対処法

- 擬似二項分布を誤差分布にする
- 混合効果ロジスティックモデル

誤差分布に二項分布属でばらつきが広い分布を持ってくるか、モデルの中に、**独立変数で説明できず、かつ誤差分布からあふれだすばらつき**を処理する項（混合効果）を組み込んでやる。

どちらの手法も **R** なら簡単に実行できます。

Overdispersionへの対処法

混合効果モデル (Generalized Linear Mixed Model; GLMM)
は今回の範囲を超えるので、またの機会に…

擬似二項分布への当てはめ

```
> model00 <- glm(cbind(Sons, Brood - Sons) ~ MTL +  
FTL + ... FTL:mBBP, family = quasibinom, data =  
para)
```

ここを変えるだけ

二値データ

- ある個体が雄か雌か
- 一回のけんかに勝ったか負けたか
- 卵が孵化したか、しなかったか
- ある個体が成鳥か幼鳥か
- ある巣が捕食されたか、無事だったか

鳥屋の皆さんがよく出会うデータ構造

従属変数が二値データのとときに生じる問題

- 誤差が正規分布しない (ベルヌーイ分布)
 - 等分散ではない ($\sigma^2 = pq$)
 - 従属変数に上限下限がある ($y = 0$ or 1)
- ※ただし、クラスター構造は無いので、
overdispersion (過分散) は生じない

やはり普通に回帰できない

ではどうするか？

昔：変数変換？

いま：一般化線形モデル(logit link, binomial error)

ベルヌーイ分布は $n = 1$ のときの二項分布なので、
binomial errorが適用できる。