2003.11.26

生物多様性論 I: 「生態学における統計学的手法の基礎」 (と勝手に改題)

全部で 3 回講義の 3

データにあわせる統計モデリングの例

— なんでも割算すんな! —

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2003/

講釈: **久保拓弥**　kubo@ees.hokudai.ac.jp

この 3 回だけの統計学授業でやること

統計学が 何もわかってない 修士課程大学院生が対象．この善男善女どもがデータ解析に際して「統計学って，まるっきり理解不能」と遭難する確率を減少せしめるべく，この世界の おおまかな地図を提供し解説する．

- **第 1 回**: 2003.11.17 (**月**)
  「検定」の使われかたを観察しよう
  — **「検定」ってそんなにエラいのか**?
- **第 2 回**: 2003.11.19 (**水**)
  統計モデリングと推定を重視してみる
  — **理解できる統計学めざして**
- **第 3 回**: 2003.11.26 (**水**)
  データにあわせる統計モデリングの例
  — **なんでも割算すんな**!

個別的なワザより全体に共通する考えかたを— ただし内容は偏ってるよ

今日のハナシ: 「あり・なし」データの統計学

1. **二値データを割算せずに**
   「あり・なし」データ— 0 or 1，生きてる vs 死んでる，オスかメスか，繁殖してるしてない，などなど— はどうやって統計モデル化するか?
2. **汎用性のある推定方法: 最尤推定**
   最も尤もらしい推定方法について
3. **いろいろな推定結果を比較してみる**
   どういうモデルを選択すればよいか?

観察: 生態学者のデータ解析 (悪い面に着目)

ゑくせる奴隷になる

計算機に使役される日々 — そして間違い混入のもと

何でもすぐに割算する・平均する

← 今日はこのあたりを ちくちくと

割算すればケガれたデータが浄化される，という素朴な信仰

ブラックボックス統計学

「それじゃおみくじなの?」 (in ぴんく本)

ゆーい差決戦主義

統計学的ゆーい差を生態学的ゆーい差とすりかえる— あなたの問題は検定でけりがつくんですか?

今日の問題: 割算やらずに二値データ解析

「あり・なし」となるようなデータはどう扱えばよいか?

分割表の検定 (今日はとりあつかわない)

- Fisher **の正確確率検定**
- **二項検定**

粕谷さんぴんく本
第 4 章「頻度のデータ – % になおすな (できるだけ)」

分割表のモデリング

(実は下のふたつは数学的には同一)

- **対数線形モデル (一般化線形モデルのポアソン分布** family)
- **ロジスティックモデル (一般化線形モデルの二項分布** family)

一般の「整数ぶんの整数」モデリング問題

- **ロジスティックモデル(一般化線形モデル**; GLM) **が便利そうだなぁ**

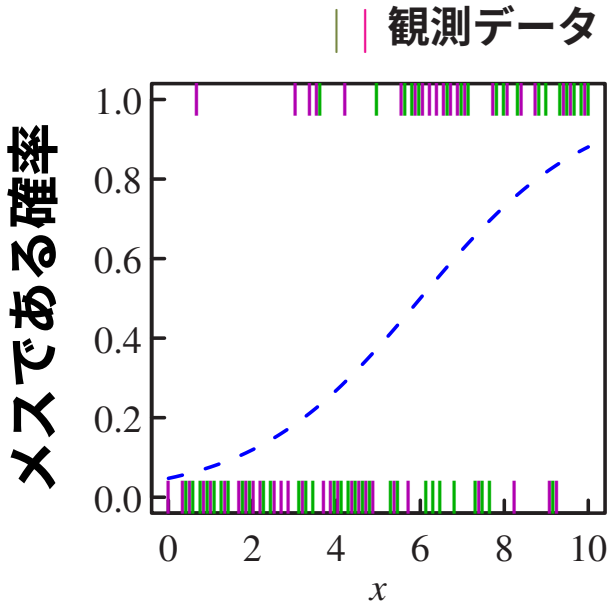
一般化線形モデル (generalized linear model, `glm()`)

- **指数関数族に属する確率分布あれこれ (正規分布，二項分布，ポアソン分布，$\cdots$) で説明されるばらつきのデータに適用できる**
- link **関数を指定できる**
- **独立変数は何でもよい: 連続変数，名義変数，順序変数**
- **パラメーターは線形に結合していなくてはならない (線形モデル)**

$$\mathrm{link}(\mu(\mathbf{x})) = \beta_0 \cdot 1 + \beta_1 x_1 + \beta_2 x_2 + \cdots = \sum_i \beta_i x_i$$

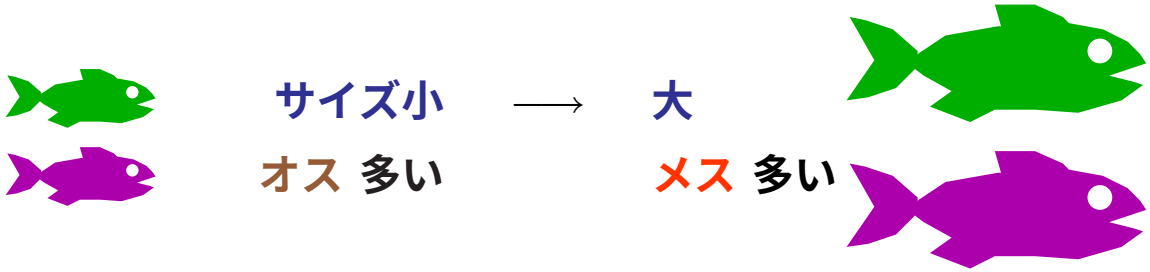
統計ソフトウェア `R` **では** `glm()` **関数で簡単に推定計算ができる**

架空生物の観測データ: 性転換するサカナ



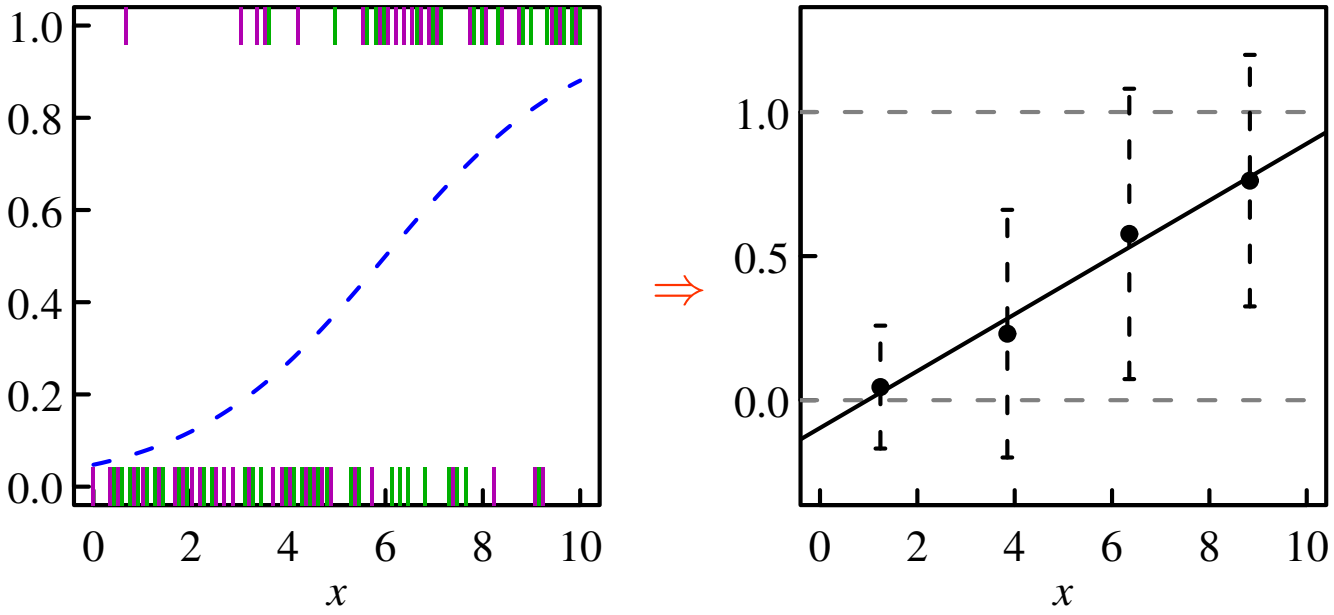
[“神” の立場で知ってるコト]

- メスである確率は 青破線で示されているように上昇する
- この確率は，魚の色 (緑 と 紫) とは 無関係
- サンプリングに少し偏りあり



青破線(つまり真のモデル・母集団) を推定したい

(よく見かける) よろしくない解析の一例



1. **てきとーにサイズ** (x) **の「区画」を取る** (**上の例だと** 4 **区画**)
2. **区画ごとに縦横の平均値など計算**; $\{0, 1\}$ **データを割算値に**
3. **統計ソフトウェアにほうりこむ**
   (**直線回帰する** or **「分散分析」する** or **「検定」&多重比較する**)

2003.11.26

なぜよろしくないか? データの特徴を無視

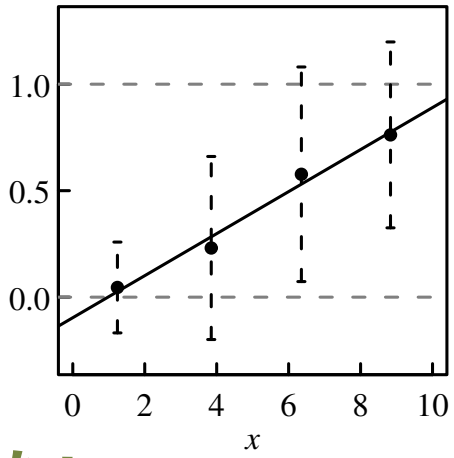
区画はてきとー

区画のとりかたで結果は変わる

割算すると情報が失われる

1 / 2 と 100 / 200 は違う!

— 十円玉なげの例で考えてみよ



等分散でもなければ正規分布でもない

ということで直線回帰も分散分析も使えん— さらに，いわば母分散が異なる状況なので，ノンパラメトリック検定のたぐいもだめ

何を推定してるのだろうか?

メスになる確率がマイナスになったり，1 をこえたりするモデルってのは……? (変数変換すればいいって? そのワザは呪われてる)

あなたのデータにぴったりの確率分布はコレ!

何でもかんでも変数変換しない・データにあわせて分布を選んで推定

— 選びかたの三つのポイント —

1. **説明したい量は離散か連続か?**
   - **離散**: { **生きてる**, **死んでる** }, **カウントデータ**, $\cdots$
   - **連続**: $\{0.56,\ 1.33,\ 12.4,\ 9.84,\ \cdots\}$, $\cdots$
2. **説明した量の範囲は?**
   - $\{0, 1, \cdots, N\}$, $\{0, 1, \cdots, \infty\}$, $[y_{\min}, y_{\max}]$, $[-\infty, \infty]$, $\cdots$
3. **説明したい量の分散 (ばらつき) と平均の関係は?**
   - **分散** $\approx$ **定数**, **分散** $\approx$ **平均**, **分散** $\propto$ **平均**, **分散** $\propto$ **平均** ^{n}, $\cdots$

確率分布は二項分布，まずはとても単純化

正確にはベルヌーイ分布 (個体を区別してるから)
……しかし以降の計算にはほとんど影響ナシ

- **サカナの色を無視する (あとで検討する)**
- **サカナのサイズを無視する (あとで検討する)**
- **オス・メスだけを見る**

単純化された観測データ:

オス　メス　メス　オス　メス

問: あるサカナがメスである確率 p を推定せよ

確率分布を推定する方法たちの階層性

なぜ一般化線形モデルを? — より広い世界を統一的に扱えるから

[**最尤推定法**で扱えるモデル]
何でもいいから確率分布があるモデル
一般化線形混合モデルなどなど

[**一般化線形モデル** (GLM)]
指数関数族の確率分布 + 線形モデル
ロジスティック回帰，ポアソン回帰などなど

[**最小二乗法**的に考えるモデル]
等分散正規分布 + 線形モデル
直線回帰，いわゆる「分散分析」などなど

最尤 (最大尤度) 推定法で考えてみる

メスである確率 p という統計モデル のもとで

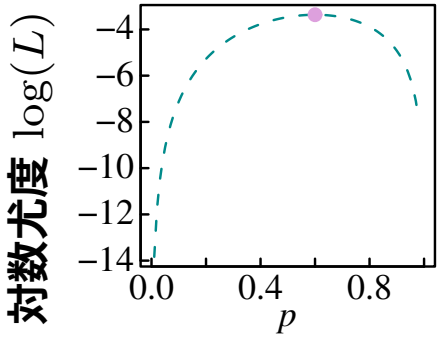
観測データ:

オス　メス　メス　オス　メス

が観測される確率は? (尤度(ゆーど))

尤度 $L = (1-p)\cdot p\cdot p\cdot (1-p)\cdot p = p^3(1-p)^2$ **尤度方程式**

$\log(L) = 3\log(p) + 2\log(1-p)$ **対数尤度方程式**



最尤推定値 $\hat{p} = 0.6$ **となる.**

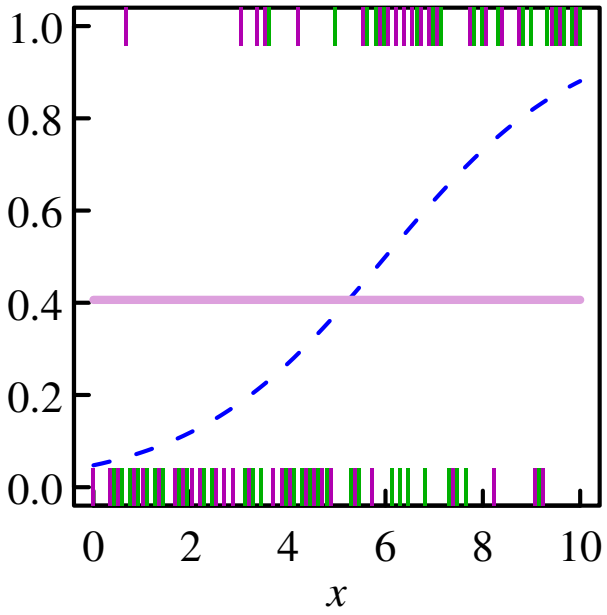
これは $\frac{\partial \log(L)}{\partial p} = 0$ **となる点を探してもよい.**

$\hat{p} = \frac{3}{3+2} = 0.6$ **が最尤推定値となる** $(0 < \hat{p} < 1)$**.**

(なんだ，割算値ぢゃん)

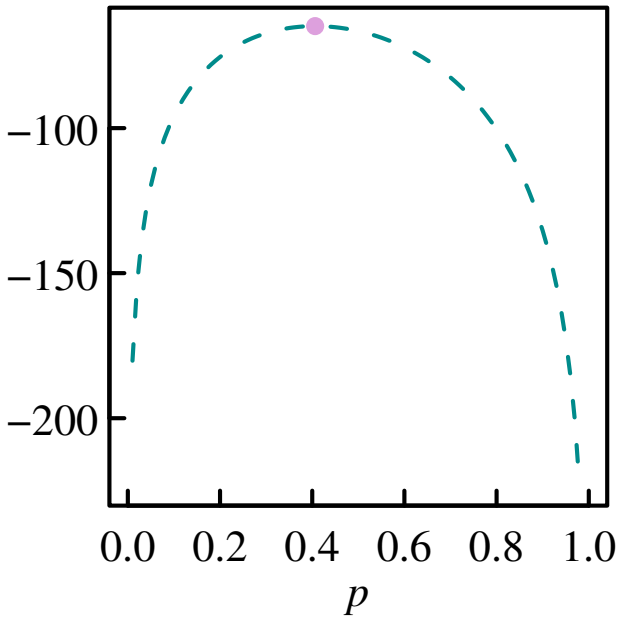
最尤推定の結果: サイズに依存しないモデル

データから $\hat{p}$ を推定する



$\Leftrightarrow$

対数尤度と p



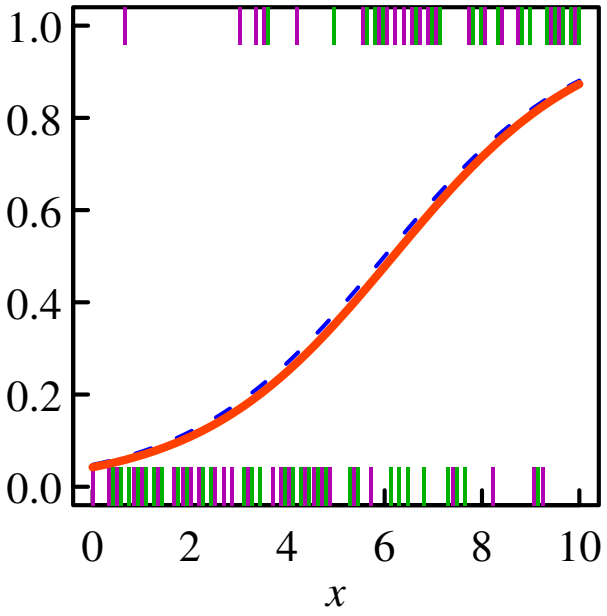
(**対数尤度の最大値はゼロ**)

R **の** `glm()`(**一般化線形モデルの計算**) **では**

```
glm(y ~ 1, family = binomial(logit) data = ...)
```

次に「サイズに依存するモデル」で最尤推定

赤線: **推定されたモデル**



最尤推定法の利点はモデルが複雑 (ここではサイズ x に依存) になった場合であっても，まったく同様に推定計算ができるところにある．

尤度方程式:

$$L = \prod_{\{\text{Female}\}} p(x_i) \prod_{\{\text{Male}\}} (1 - p(x_i))$$

ただし $p(x) = \dfrac{1}{1 + \exp(-(a + bx))}$.

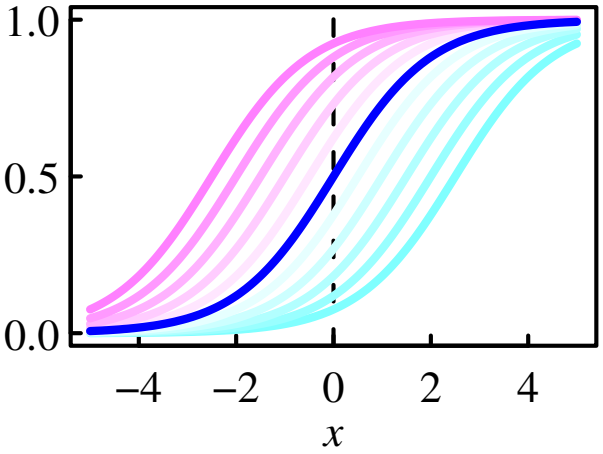
R **の** `glm()`(**一般化線形モデルの計算**) **では**

```
glm(y ~ 1 + x, family = binomial(logit) data = ...)
```

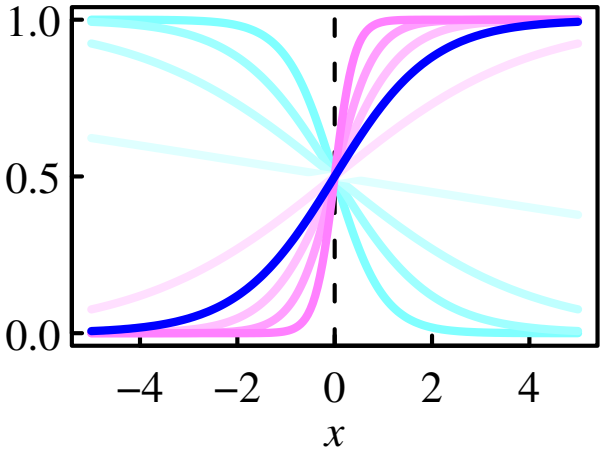
そのロジスティック曲線って何なの?

$$p(x) = \frac{1}{1 + \exp(-(a + bx))} \qquad (\exp(z) = \mathrm{e}^z \text{ のこと})$$

a だけ変化させる



b だけ変化させる



どうしてもロジスティック曲線でなきゃならん，という理由は何も**ない**— 簡単かつ便利なんで，なンとなく惰性で使ってるだけ

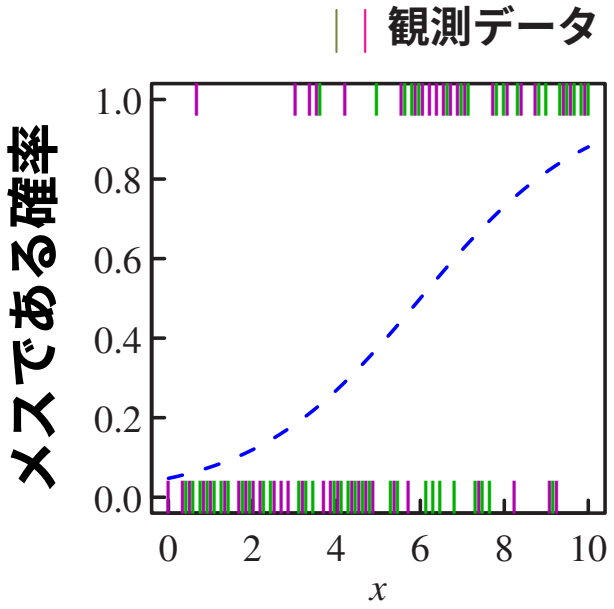
ここまで: 確率分布を考えて最尤推定

自力で最尤推定法を構成できないあいだは，一般化線形モデル `glm()` で使える確率分布から検討するのがよいだろう

	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	`rbinom()`	`glm(family = binomial)`
	二項分布	`rbinom()`	`glm(family = binomial)`
	ポアソン分布	`rpois()`	`glm(family = poisson)`
	負の二項分布	`rnbinom()`	`glm.nb()`
(連続)	ガンマ分布	`rgamma()`	`glm(family = gamma)`
	正規分布	`rnorm()`	`glm(family = gaussian)`

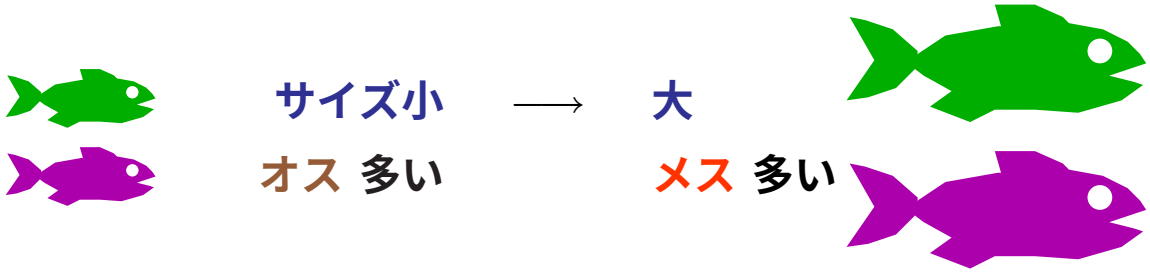
(前回の授業を参照)

性比には無関係な「体色」にだまされてみる



[“神” の立場で知ってるコト]

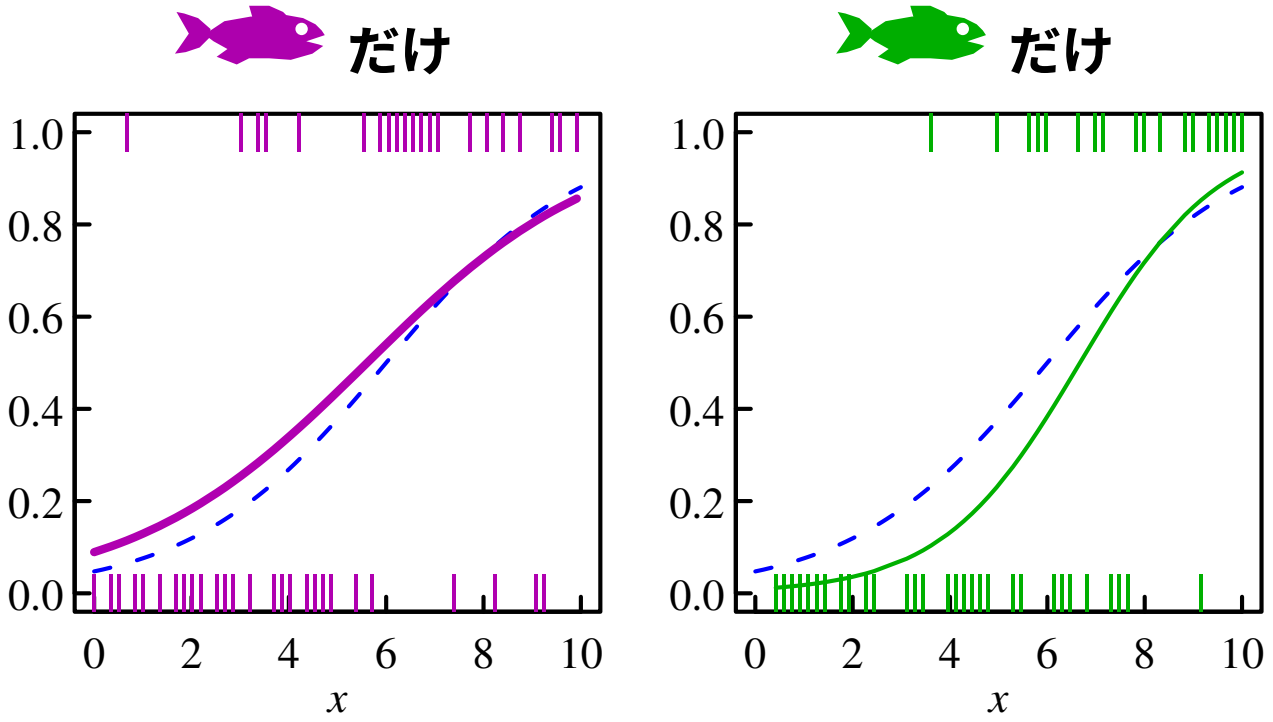
- メスである確率は 青破線で示されているように上昇する
- この確率は，魚の色 (緑 と 紫) とは 無関係
- サンプリングに少し偏りあり



サカナの色によってメスになる確率が変わる，とだまされてみよう

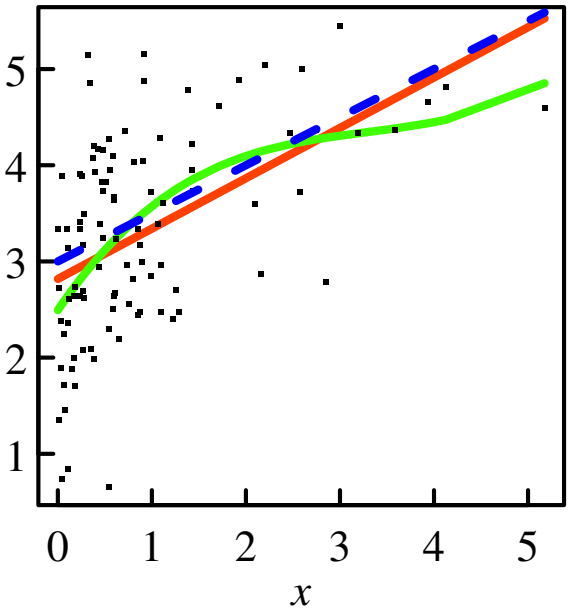
えっ? サカナの体色でオス・メスが変わる?

ためしに `glm()` でパラメーター推定してみると……



「体色」(性比にまったく無関係) ごとにわけて，個別に推定したほうが観測データ (標本) へのあてはまりは良くなってるよーな……

ちょっと別の問題で考えてみる: 単純 vs 複雑



[真のモデル]

青破線: $y = a + bx$

を平均とする等分散な正規分布

[推定に使ったモデル]

赤線: $y = a + bx$ (**単純**)

緑線: $y = a + b_1 x + b_2 x^2 + b_3 x^3$ (**複雑**)

- **「複雑」なモデルのほうが「雑音」をひろってしまいやすい**
- **われわれがやりたいことは，標本 (黒い点々) に最も「あてはまる」線を生成したいわけ**ではなく**，標本を生み出した母集団 (つまり青破線) の推定である**

モデル選択: あてはまりと複雑さのバランス

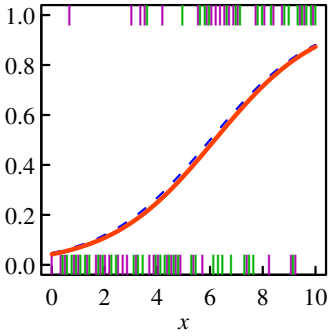
- じつは，どんな場合でも「小分けにしてあてはめ」「モデル複雑化」やったほうが全体のあてはまりが良くなる

「あてはまりのよさ」は最大化対数尤度でわかる:

$-2\times$ (最大化対数尤度) $=$ ← これが小さいほど「あてはまりが良い」

97.7	95.9	97.1	94.4

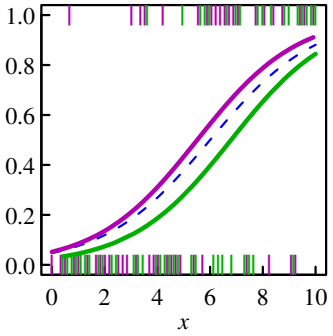
色による区別なしモデル



$a + bx$

2 パラメーター

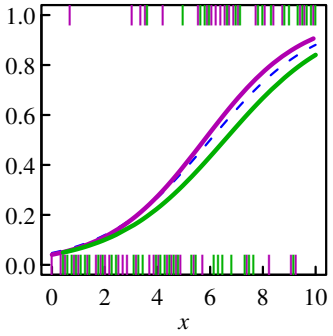
定数項だけ違うモデル



$\{a, a\} + bx$

3 パラメーター

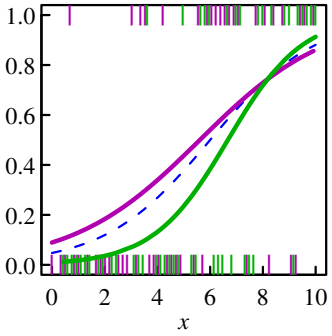
サイズ依存項だけ違うモデル



$a + \{b, b\}x$

3 パラメーター

色ごとに推定したモデル



$\{a, a\} + \{b, b\}x$

4 パラメーター

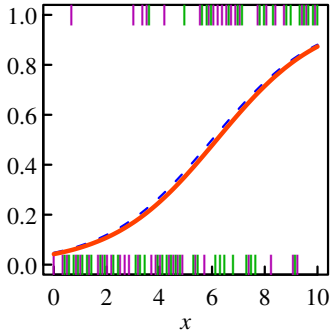
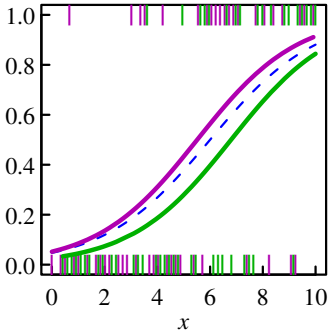
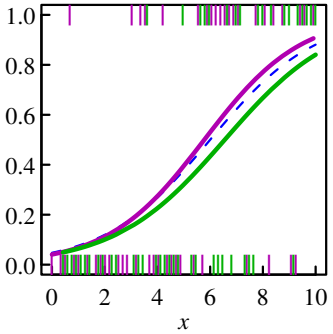
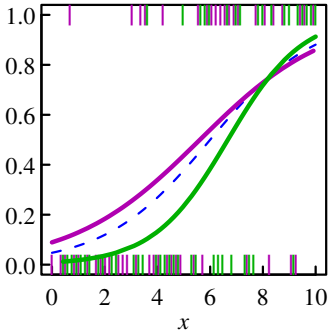
モデル選択基準: ここでは AIC 使ってみる

Akaike's Information Criteria (**赤池の情報量基準**)

これが小さいほど「良い」モデル

$$\mathrm{AIC} = -2(\textbf{最大化対数尤度}) + 2\,(\textbf{パラメーター数})$$

あてはまりぐあい (良い)　モデルの複雑さ (悪い)

色による区別なしモデル	定数項だけ違うモデル	サイズ依存項だけ違うモデル	色ごとに推定したモデル
			
$a + bx$	$\{a, a\} + bx$	$a + \{b, b\}x$	$\{a, a\} + \{b, b\}x$
2 パラメーター	3 パラメーター	3 パラメーター	4 パラメーター
$\mathrm{AIC} =$ **101.7**	101.9	103.1	102.4

単純なわりにあてはまってるモデルは AIC が低い

ここまで: モデル選択について

- 得られた標本へのあてはまりの良いモデルは「良い」
- 一方で，複雑すぎるモデルは良くない— 単なる「雑音」ひろってるだけなんで
- 母集団すなわちもともとの統計モデル (青破線) に「近い」モデルが選ばれる— こういう事例に関して AIC はうまく働いてるようだ

統計ソフトウェア R では stepAIC() 関数で簡単に AIC によるモデル選択ができる

本日のまとめ

1. **割算すんな!**
   整数ぶんの整数，という問題は**二項分布**なモデルでまずは解析してみる
2. **統計学の基本，最尤推定法を理解しよう**
   そのためには「この現象はどういう**確率分布**で説明されるか」をよく考える
3. **いろいろある中から「良い」モデルを選択**
   「**あてはまり**良く」「パラメーター数の少ない (つまり**単純**な)」モデルを選択する AIC などなど