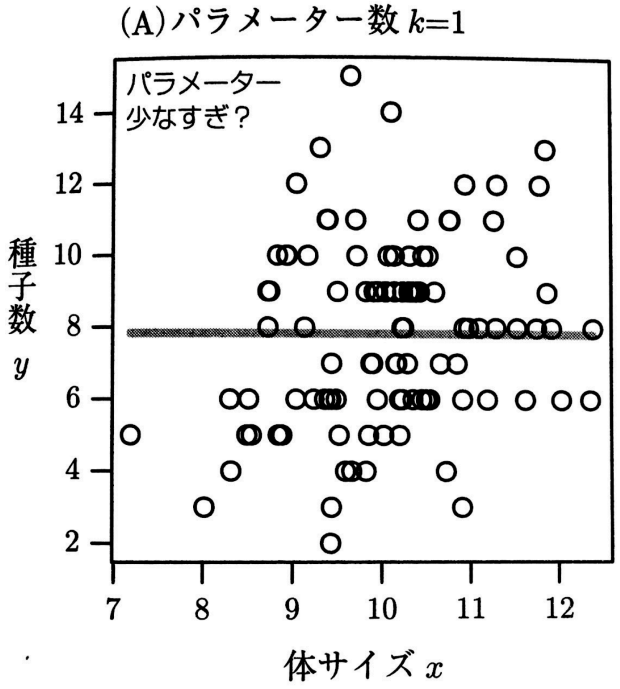
第4章:GLMのモデル選択(前半)

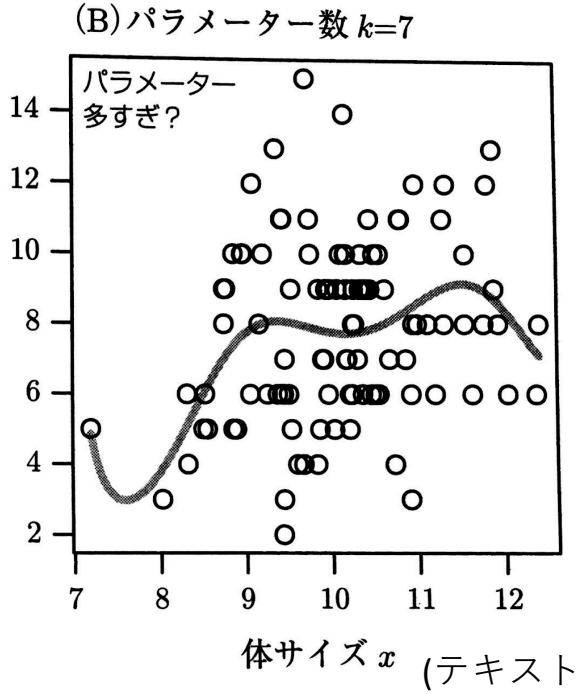
太田研究室 学部4年 和田

「良い」モデルは最大対数尤度のみでは決定されない

パラメーター数(k)を多くすればする分、最大対数尤度は大きくなる



$$\log\lambda = \beta_1$$
$$k = 1$$



$$\log\lambda = \beta_1 + \beta_2 x + \ldots + \beta_7 x^6$$
$$k = 6$$

(テキストより)

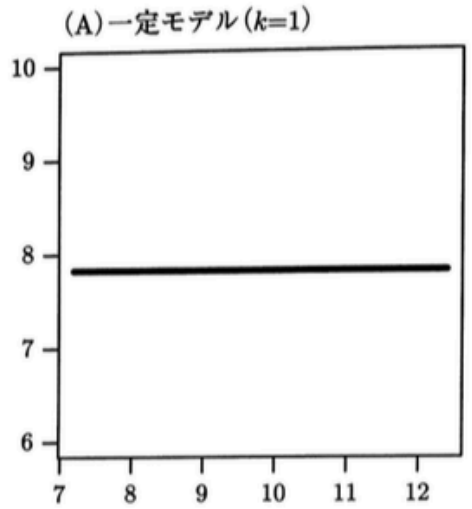
こちらの方がモデルとして優れている
……とは限らない！

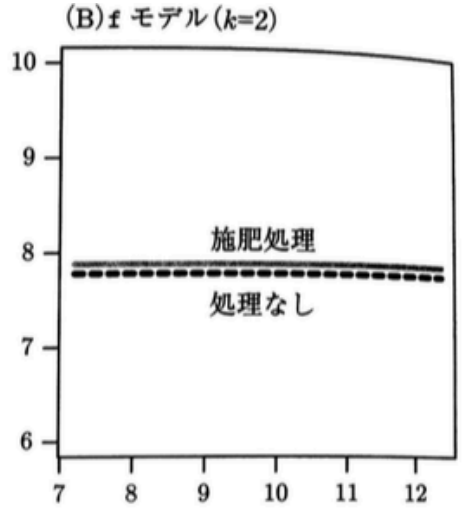
【理由】
・計算処理
・実際の現象との乖離

最大対数尤度(=観測データへの当てはまりの良さ)以外の、新しいモデルの選択基準：
「そのモデルは良い予測をするのか？」
→「AIC」で判断可能

一つのデータに対し、考慮する説明変数のパターン(=候補となるモデル)はたくさんある

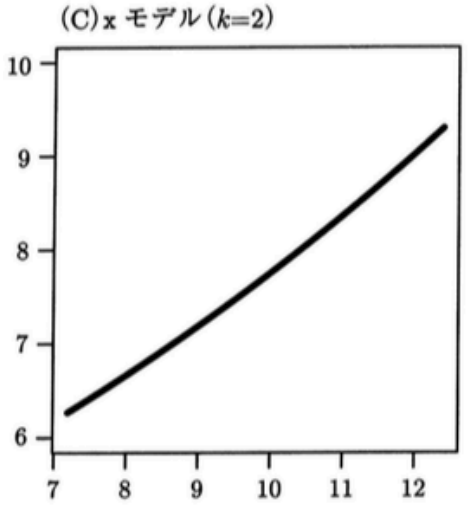
体のサイズx_iも施肥の有無も、種子の量y_iに影響しない

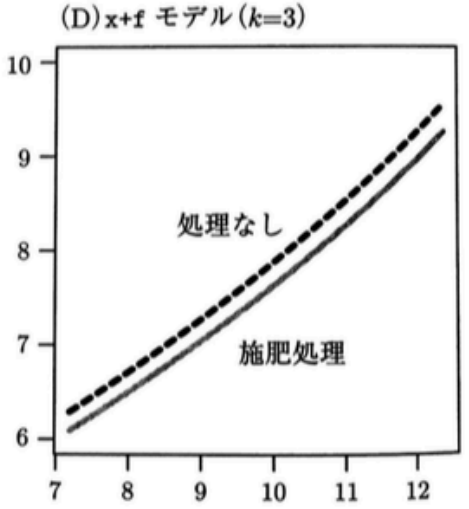




施肥の有無f_iのみが種子の量y_iに影響する

体のサイズx_iのみが種子の量y_iに影響する





施肥の有無f_iと、体のサイズx_iの両方が、種子の量y_iに影響する

(テキストより)

一つのデータに対し、考慮する説明変数のパターン(=候補となるモデル)はたくさんある

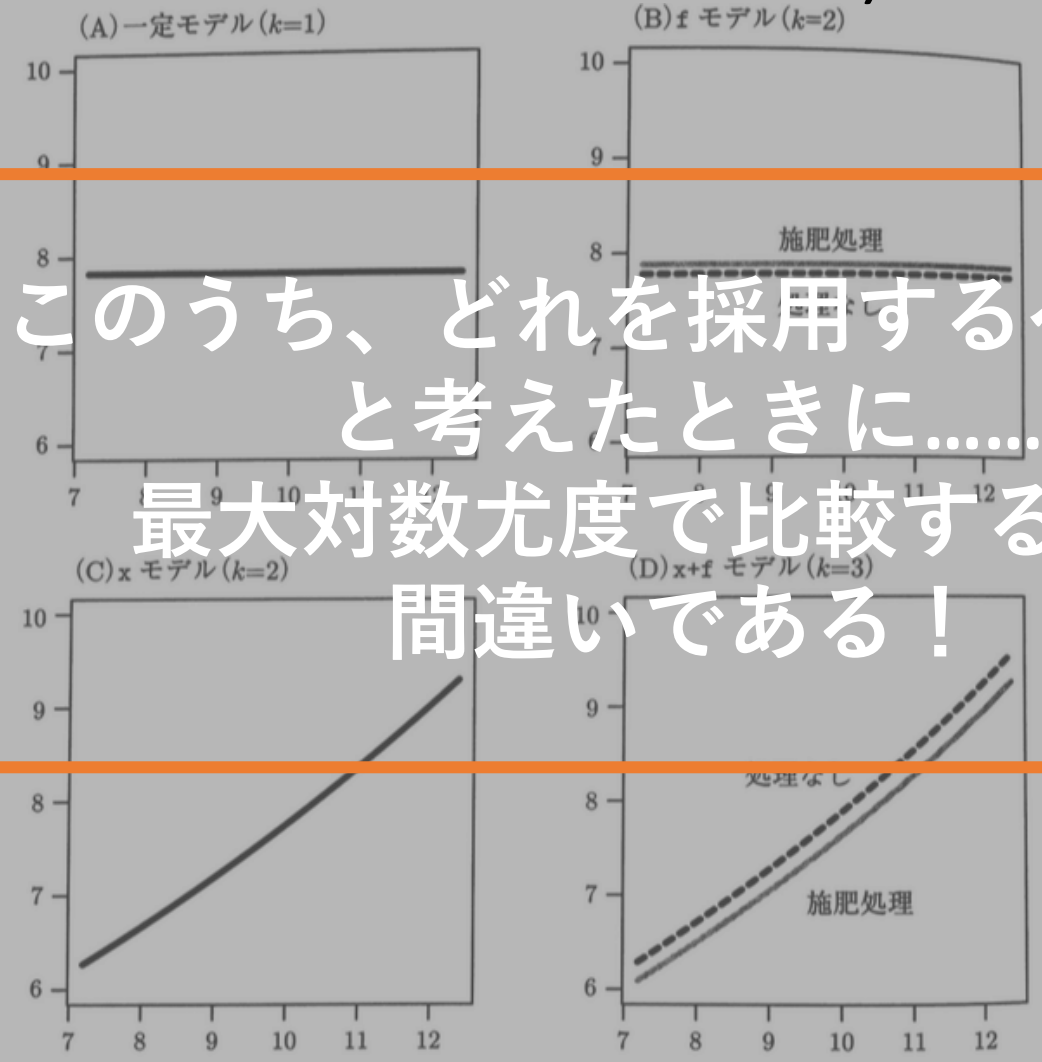


図 4.2　第 3 章の例題データを説明する 4 種類のポアソン回帰

体のサイズx_iも施肥の有無も、種子の量y_iに影響しない

施肥の有無f_iのみが種子の量y_iに影響する

体のサイズx_iのみが種子の量y_iに影響する

施肥の有無f_iと、体のサイズx_iの両方が、種子の量y_iに影響する

このうち、どれを採用するべきか？
と考えたときに……
最大対数尤度で比較するのは
間違いである！

モデルのデータへのあてはまりの悪さ「逸脱度」は、最大対数尤度の変形

- 逸脱度＝Deviance
- 統計モデルの、データへの「あてはまりの悪さ」の指標

$$D = -2\,logL^{*}$$

$logL(\{\beta_j\})$を$logL$、
その最大対数尤度を$logL^{*}$と表記

glm()コマンドの出力結果に表示

名前	In English	定義
逸脱度 (D)	Deviance	$-2\,logL^{*}$
最小の逸脱度	Minimum deviance	フルモデル(後述)をあてはめたときのD
残差逸脱度	Residual deviance	D – 最小のD
最大の逸脱度	Maximum deviance	Nullモデル(後述)をあてはめたときのD
Null 逸脱度	Null deviance	最大のD – 最小のD

フルモデル、Nullモデルはそれぞれ、パラメーター数を最大、最小(1)にした場合のモデルである①

「**フルモデル**」(full model)　…　最もあてはまりがいいモデル

- 個々のデータに、一対一対応でパラメーターλが定まっている
  - 100個のデータがあれば100個のλを定めている

例：$y_i = \{6,6,6,12, \ldots\}$のとき、
$i \in \{1,2,3\}$の y_i は6なので、$\{\lambda_1, \lambda_2, \lambda_3\} = \{6,6,6\}$
$i = 4$のy_4は12なので、$\lambda_4 = 12$
…(以後同上)

フルモデルを当てはめた時の
逸脱度＝最小のD(minimum deviance)

- (同じ回帰で)他のどのモデルを使った時よりも、必然的に対数尤度は最大、逸脱度は最小になる
- 「現象を説明しうる理想のモデルを考えている」のではなく、「現在のデータ(のみ)にモデルを近づけている」ので、モデルとしての価値はない

フルモデル、Nullモデルはそれぞれ、パラメーター数を最大、最小(1)にした場合のモデルである②

「**Null モデル**」(Null model)　…　最もあてはまりが悪いモデル

- パラメーター数が1
  - つまり、この文脈においては$\lambda = e^{\beta_1}$
- パラメーターは、全ての説明変数から完全に独立である
- (同じ回帰で)他のどのモデルを使った時よりも、必然的に対数尤度は最小、逸脱度は最大になる

Null モデルを当てはめた場合の逸脱度
＝最大のD (Maximum deviance)

種々の逸脱度の関係性は以下

(あてはまりが悪い)

最大のD

一つのデータ
に対し固定

Nullモデル

結果を見てみたい
任意のモデル

モデル①のD

モデル①

Null deviance(Null逸脱度)

そのモデルの当てはまりの良さ

Residual deviance
(残差逸脱度)

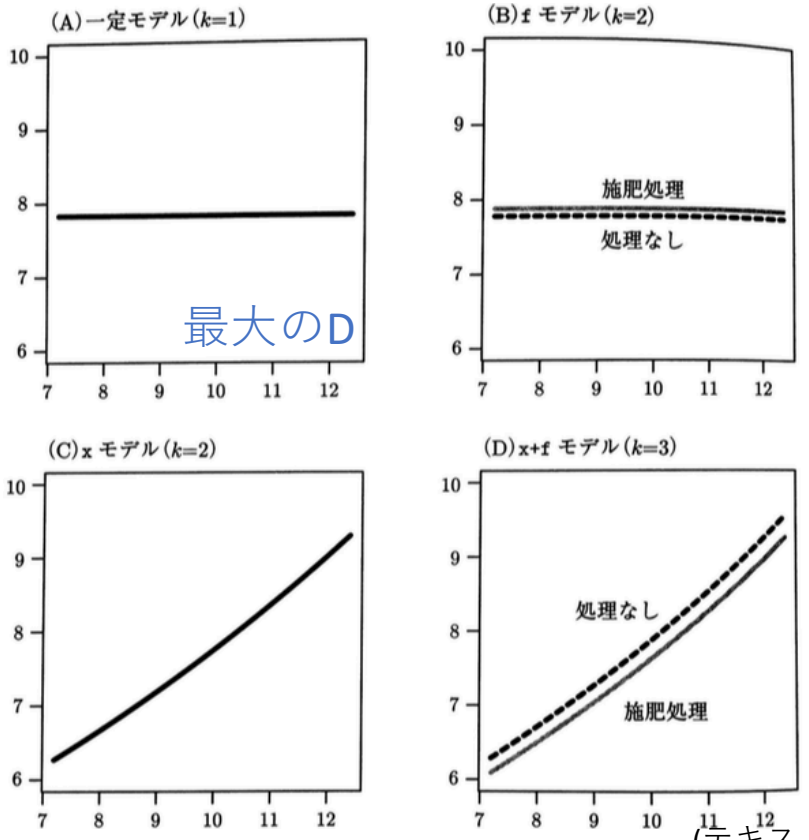
モデル①の
残差逸脱度

Decianve=逸脱度
$-2\ logL$

最小のD

フルモデル

(あてはまりがいい)



パラメーター数が増えるほど、残差逸脱度は減る

(テキストより)

AICの比較により「予測の良さ」を重視したモデル選択を行うことができる

AIC (Akaike's information criiterion)

- 「モデル選択基準」(model selection criterion)の一つ
- 予測の良さを重視する (あてはまりの良さ、ではない)
- 小さい方が「良い」モデル

$$AIC = -2\{(\text{最大対数尤度}) - (\text{最尤推定したパラメーター数})\}$$
$$= -2(logL^{*} - k)$$
$$= D + 2k$$

残差逸脱度とパラメーター数が小さい時が「良い」モデル

(AICがモデル選択基準として有効である理由については次回)