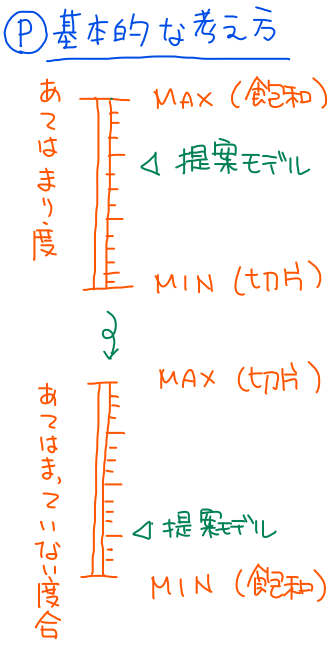



## 📖 ノート３　評価１：モデルのデータへのあてはまりの良さ

### ○ 考えたいこと

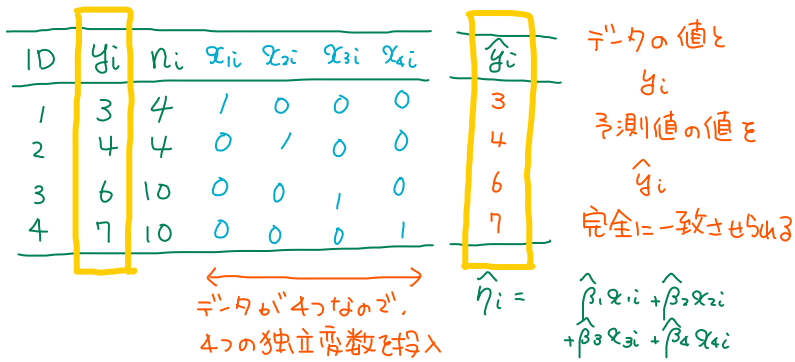
重回帰では、予測値$\hat{y}_i$と観測値$y_i$の差である残差やこの二つの相関指標である決定係数を扱った。

⇒ 同じように、ロジスティック回帰における「モデルのデータへの当てはまりの指標」を作ろう！

### （１）比較の基準となるモデルを作る

#### ① 飽和モデル Saturated model　⇒あてはまり度 MAX

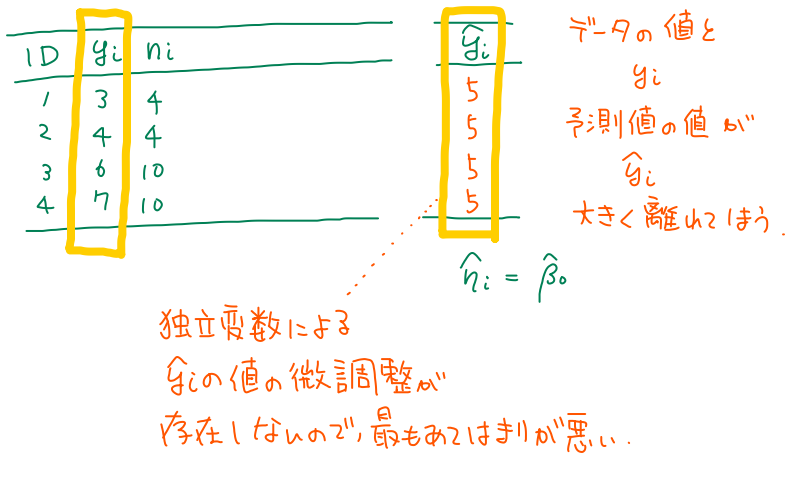
これは、推定されうるパラメータの最大個数を含んだモデル。最大モデル（maximal）、フルモデル（full model）。



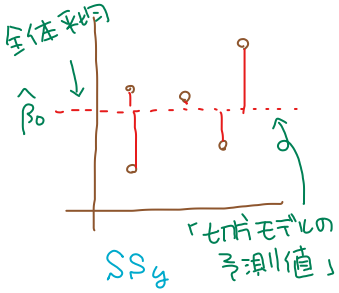

#### ② 切片モデル Null model　⇒あてはまり度 MIN

これは、上記で「独立変数を含まないモデル」と呼んでいた切片しか含まないモデル。一定／最小モデル。

⇒ どんな独立変数も確率に影響を与えないという強い仮定を持っていることになる。



Ⓟ データ$y$の平方和

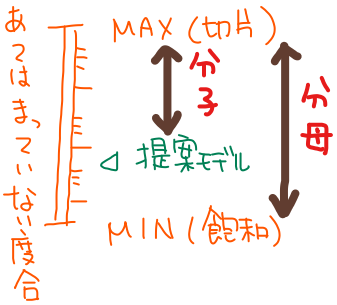



＝「全体の平均」から各点がどれだけばらついているか．

＝「切片モデルの予測値」から各点がどれだけばらついているか

＝ $SS_{\text{切片モデル}}$

Ⓟ あてはまらなさの比較




（2）決定係数$R^2$（復習）

「最小二乗法」という視点から提案モデルがデータにフィットしている度合いを測るための指標として使われる。

① これまでの定義

$$R^2 = \frac{SS_{\hat{y}}}{SS_y}$$

$$= \frac{SS_y - SS_e}{SS_y}$$

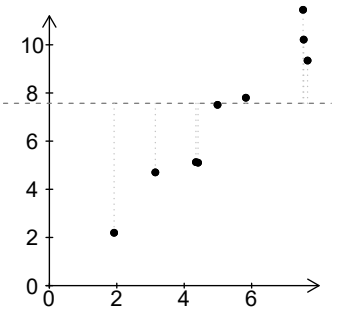
② 別の見方

$$R^2 = \frac{SS_{\text{切片モデル}} - SS_{\text{提案モデル}}}{SS_{\text{切片モデル}} - 0}$$

$$= \frac{SS_{\text{切片モデル}} - SS_{\text{提案モデル}}}{SS_{\text{切片モデル}} - SS_{\text{飽和モデル}}}$$

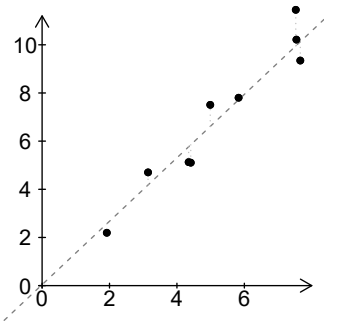
**残差平方和は「あてはまりの悪さ」の指標**

① データ全体のばらつき（＝切片モデルの残差）




$\hat{y}_i = \hat{\beta}_0$

② 残差のばらつき（＝提案モデルの残差）




$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$

**残差という概念の拡張**

「最尤推定法」という視点から提案モデルがデータにフィットしている度合いを測るための指標を次のような形で作りたい。

$$R^2 = \frac{\ell^*_{\text{切片モデル}} - \ell^*_{\text{提案モデル}}}{\ell^*_{\text{切片モデル}} - \ell^*_{\text{飽和モデル}}}$$

## （3）残差逸脱度 Residual Deviance

これは、最大対数尤度の点で提案モデルが飽和モデルに対しどの程度「あてはまりの悪さ」を持つのかを表す指標。

飽和モデルにおける あてはまりの良さ　　提案モデルにおける あてはまりの良さ

$$D_{\text{提案}} = 2 \times [\ell^*(\text{飽和モデル}) - \ell^*(\text{提案モデル})]$$

飽和モデルにおける 最大対数尤度　　提案モデルにおける 最大対数尤度

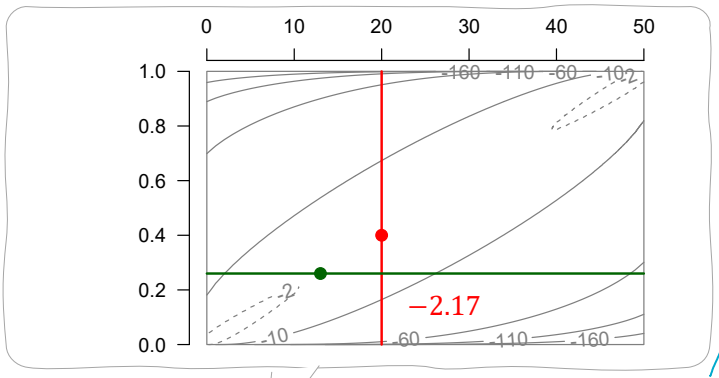
### ① 最大対数尤度

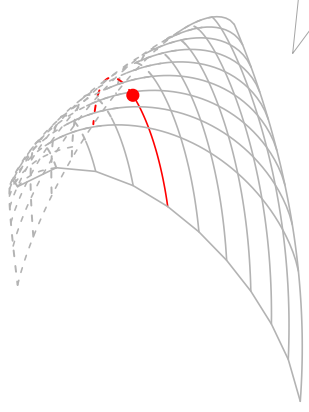
これは、対数尤度関数の引数$\beta_0, \beta_1, \ldots$に最尤推定値$\hat{\beta}_0, \hat{\beta}_1, \ldots$を代入した値のこと。

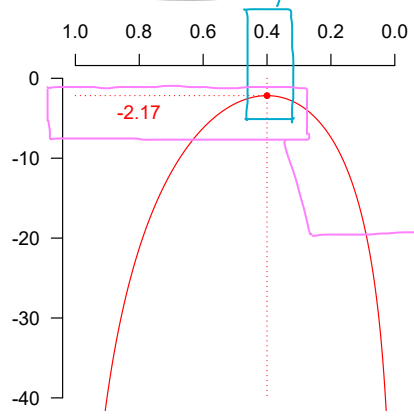
$$\begin{aligned} \ell^* &= \ell(y|\hat{\beta}_0, \hat{\beta}_1, \ldots, n_i) \\ &= \log L(\boldsymbol{y}|\hat{\beta}_0, \hat{\beta}_1, n_i) \end{aligned}$$

### 最大対数尤度の直感的な意味合い

最尤推定値のもとでデータ$\boldsymbol{y}$が得られる確率を表している。値が高いならデータとモデルの整合性が高いことを意味するため「モデルの当てはまりの良さ」の指標として使われる。



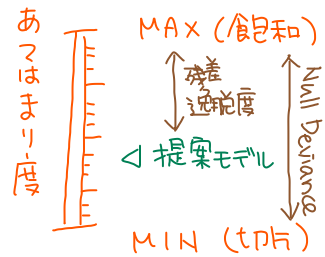






最尤推定法は，対数尤度関数がMAXになるパラメータの値（例：π）を求めていた．

最大対数尤度は，最尤推定値（$\hat{\pi}_{MLE}$）における対数尤度関数の値．

### Ⓟ残差逸脱度




### Ⓟ決定係数とのつながり

$$\begin{aligned} &1 - \frac{\text{残差逸脱度}}{\text{Null Deviance}} \\ &= 1 - \frac{\ell^*_{飽} - \ell^*_{提}}{\ell^*_{飽} - \ell^*_{切}} \\ &= \frac{\ell^*_{飽} - \ell^*_{切} - (\ell^*_{飽} - \ell^*_{提})}{\ell^*_{飽} - \ell^*_{切}} \\ &= \frac{\ell^*_{提} - \ell^*_{切}}{\ell^*_{飽} - \ell^*_{切}} \times \frac{-1}{-1} \\ &= \frac{\ell^*_{切} - \ell^*_{提}}{\ell^*_{切} - \ell^*_{飽}} \end{aligned}$$

② 最大値：切片モデルの残差逸脱度 Null Deviance

これは、切片モデルにおける残差逸脱度であり、考えられるモデルたちの中の残差逸脱度の最大値のこと。

$$D_{null} = 2 \times [\ell^*(\text{飽和モデル}) - \ell^*(\text{切片モデル})]$$

③ 最小値：飽和モデルの残差逸脱度

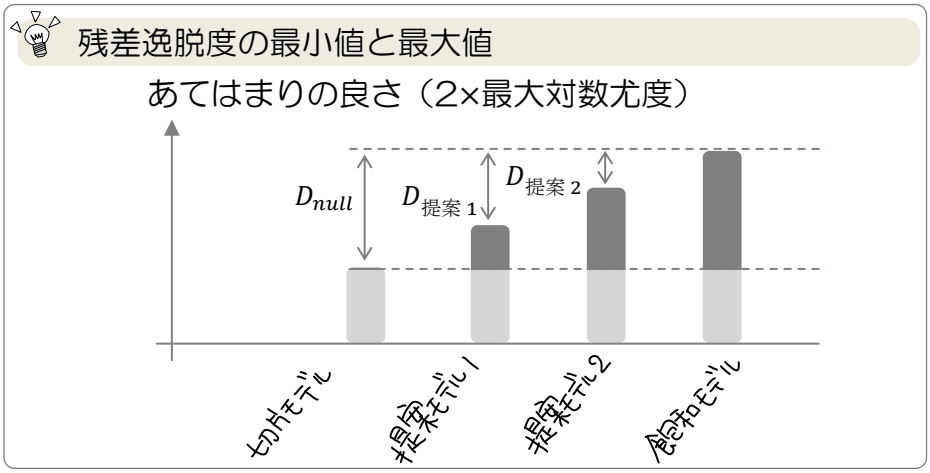
これは、飽和モデルにおける残差逸脱度であり、考えられるモデルたちの中の残差逸脱度の最小値のこと。

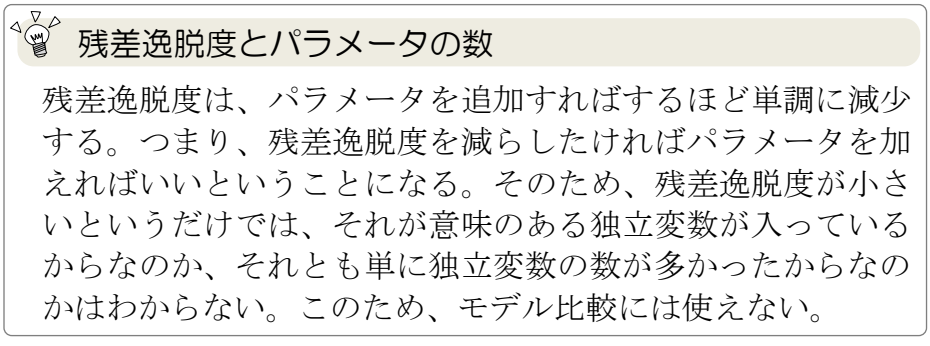
$$\begin{aligned} D_{sat} &= 2 \times [\ell^*(\text{飽和モデル}) - \ell^*(\text{飽和モデル})] \\ &= 0 \end{aligned}$$

※ 逸脱度 Deviance

これは、飽和モデルとの比較で相対化されていない「あてはまりの悪さ」の指標。

$$Dev_{\text{提案}} = -2 \times \ell^*(\text{提案モデル})$$



## 📖 ノート４　評価２：係数の推定値の正確さの評価

○ 考えたいこと

重回帰では「点推定」の他に「検定」と「区間推定」についても考察対象に据えていた。

⇒ロジスティック回帰の場合はどうなのだろう？

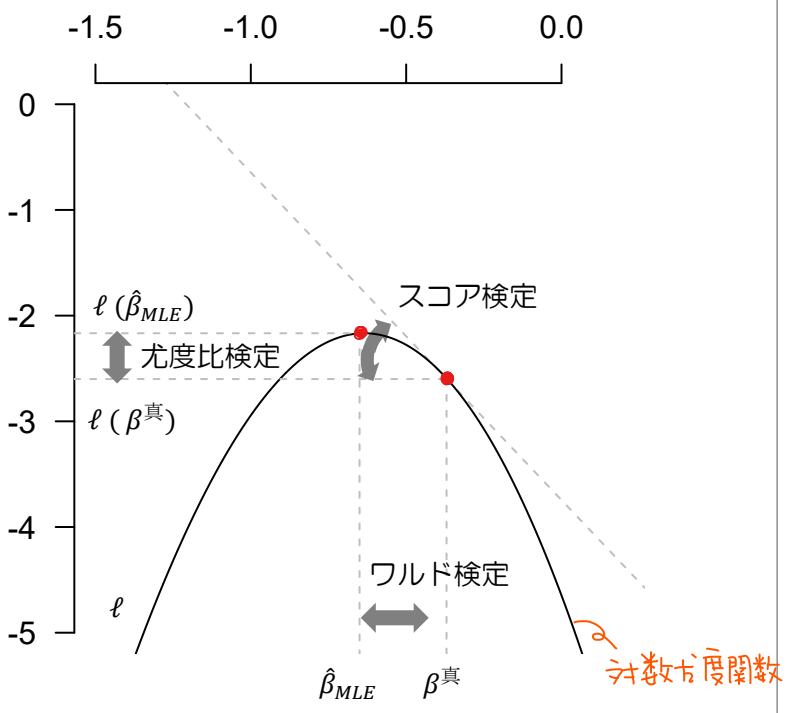
■ （対数）尤度関数に基づく三つの仮説検定

尤度比検定： H0 のもと検定統計量を $\chi^2$分布 で近似。

ワルド検定： H0 のもと検定統計量を 正規分布 で近似。

スコア検定： H0 のもと検定統計量を 正規分布 で近似。

（対数）尤度関数に基づく三つの検定



## （1）尤度比検定 Likelihood Ratio Test

これは、今回の標本で計算された二つのモデルの逸脱度の差が有意な差といえるのかどうかを検討する検定。

### ① 帰無仮説と対立仮説

帰無仮説（モデル１）：$\widehat{\beta_{MLE}} = \beta^{真}$

対立仮説（モデル２）：「$\widehat{\beta_{MLE}} = \beta^{真}$」ではない。

### ② 検定統計量：比較したいモデルの逸脱度の差

$$D_{2,1} = 2 \times [\ell^*(\text{モデル } 2) - \ell^*(\text{モデル } 1)]$$

### ③ 検定統計量が従う標本分布

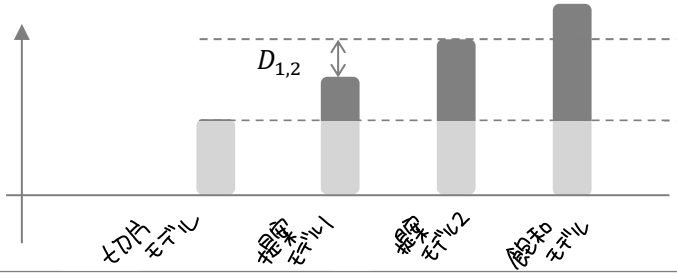
モデル１、２のパラメータ数を$p_1$、$p_2$とするとサンプルサイズが大きいとき検定統計量は $\chi^2(p_2 - p_1)$ に従う。

$$D_{2,1} \sim \chi^2(p_2 - p_1)$$

※ ここでは、ある一つの変数のみに焦点を当てているので$p_2 - p_1 = 1$となる。

**検定統計量**

あてはまりの良さ（2×最大対数尤度）



## （2）ワルド検定 Wald Test

これは、標準化された最尤推定量が漸近的に正規分布に従うということに注目して行う仮説検定。

### ① 帰無仮説と対立仮説

帰無仮説（モデル１）：$\widehat{\beta_{MLE}} = \beta^{真}$

対立仮説（モデル２）：「$\widehat{\beta_{MLE}} = \beta^{真}$」ではない。

### ② 検定統計量：比較したいモデルの逸脱度の差

$$z = \frac{\hat{\beta}_{MLE} - \beta^{真}}{\hat{\sigma}_{\hat{\beta}_{MLE}}}$$

### ③ 検定統計量が従う標本分布

正規化された統計量は、$\beta^{真}$=0 で、サンプルサイズが大きいとき、標準正規分布$N(0,1)$に近似されていく。

$$z \sim N(0,1)$$

### ④ 信頼区間

「z 値が受容域に入る確率」が 95%をもとに、「標本から計算された区間が母集団パラメータの値を含む確率」が 95%になるように信頼区間を形成する。

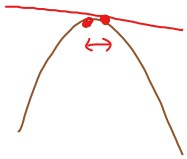
### R における実装

```
fit <- glm(y ~ x1 + x2, data = data, family = binomial(link = "logit"))
summary(fit)
```

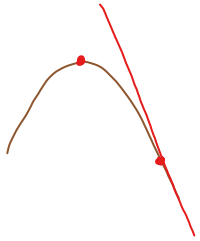
| | 推定値 | 標準誤差 | z 値 | p 値 | |
|---|---|---|---|---|---|
| 切片 | -0.38 | 0.08 | -4.98 | 6.48e-07 | *** |
| $x_1$（1 モーラ） | 0.10 | 0.24 | 0.42 | 0.67 | |
| $x_2$（意図性） | -0.18 | 0.09 | -1.95 | 0.05 | . |

Ⓟ いろいろな状況

(ケース1) 0に近いとき



(ケース2) 0から離れているとき



## （3）†スコア検定 Score Test

これは、$\beta^{真}$における対数尤度関数の 接線の傾き（スコア関数の値） が有意に0から離れているかどうかを検討する仮説検定。

① 帰無仮説と対立仮説

帰無仮説（モデル１）：$\hat{\beta}_{MLE} = \beta^{真}$

対立仮説（モデル２）：「$\hat{\beta}_{MLE} = \beta^{真}$」ではない。

② 検定統計量：

$$z = \frac{u\left(\beta^{真}\right)}{\Im_n\left(\beta^{真}\right)}$$

③ 検定統計量が従う標本分布

正規化された統計量は、$\beta^{真}$=0 で、サンプルサイズが大きいとき、標準正規分布$N(0,1)$に近似されていく。

$$z \sim N(0,1)$$

## 📖 ノート５　評価３：想定した統計モデルの適切さの評価

### ○ 考えたいこと １：モデル選択

重回帰では、情報量基準やクロスバリデーションでモデルの性能を見た。

⇒ ロジスティック回帰でも、モデルのパフォーマンスを測る指標を作ろう！

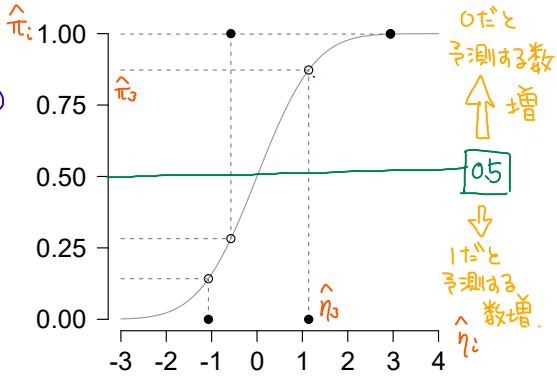
### ■ 分類表に基づくもの

#### （１）予測確率から予測値へ

##### ① ロジスティック回帰の予測するもの

ロジスティック回帰がモデル化しているのは$\hat{\pi}_i$であり$\hat{y}_i$ではない。$\hat{\pi}_i$から$\hat{y}_i$を導くルールを立てる必要がある。

| ID | $y_i$ | $\hat{\eta}_i$ | $\hat{\pi}_i$ | $\hat{y}_i$ |
|---|---|---|---|---|
| 1 | 1 | -0.58 | 0.28 | 0 |
| 2 | 1 | 2.94 | 0.99 | 1 |
| 3 | 0 | 1.13 | 0.87 | 1 |
| 4 | 0 | -1.07 | 0.14 | 0 |
| 5 | 1 | 2.17 | 0.98 | 1 |
| 6 | 0 | -0.18 | 0.42 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |




##### ② 閾値の設定

$\hat{\pi}_i$が、ある閾値$\pi_0$を超えたら$\hat{y}_i = 1$と予測するというような$\pi_0$を定める。

（例１）0.5 に設定する

（例２）標本における$y = 1$になる割合に設定する

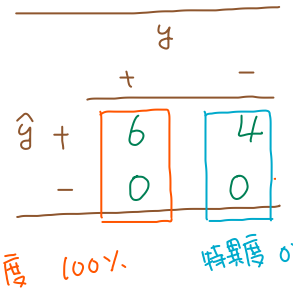
#### （２）分類表 Classification Table

テストデータを用意し、テストデータの値の予測に成功するかどうかで分類表を作成することができる。

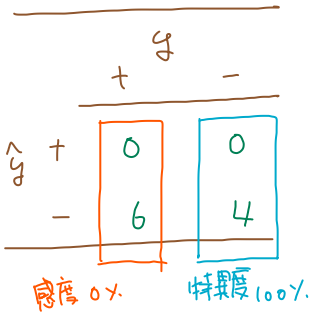
| | | 真のカテゴリー | |
|---|---|---|---|
| | | + | - |
| モデルの予測 $\hat{y}_i$ | + | 真陽性 (True positive) | 偽陽性 (False positive) |
| | - | 偽陰性 (False negative) | 真陰性 (True negative) |

## （3）分類表に基づく指標

Ⓟ 感度をMAXにする.




Ⓟ 特異度をMAXにする




### ① 感度 Sensitivity (Recall, True Positive Rate, TPR)

$$TPR = p(\hat{y} = 1|y = 1)$$

| | | 真のカテゴリー | |
|---|---|---|---|
| | | + | - |
| モデルの予測 | + | 真陽性 (True positive) | 偽陽性 (False positive) |
| | - | 偽陰性 (False negative) | 真陰性 (True negative) |

### ② 特異度 Specificity (True Negative Rate, TNR)

$$TNR = p(\hat{y} = 0|y = 0)$$

| | | 真のカテゴリー | |
|---|---|---|---|
| | | + | - |
| モデルの予測 | + | 真陽性 (True positive) | 偽陽性 (False positive) |
| | - | 偽陰性 (False negative) | 真陰性 (True negative) |

### ③ 精度 Precision (Positive Predictive Value, PPV)

$$PPV = p(y = 1|\hat{y} = 1)$$

| | | 真のカテゴリー | |
|---|---|---|---|
| | | + | - |
| モデルの予測 | + | 真陽性 (True positive) | 偽陽性 (False positive) |
| | - | 偽陰性 (False negative) | 真陰性 (True negative) |

### ④ Negative Predictive Value (NPV)

$$NPV = p(y = 0|\hat{y} = 0)$$

| | | 真のカテゴリー | |
|---|---|---|---|
| | | + | - |
| モデルの予測 | + | 真陽性 (True positive) | 偽陽性 (False positive) |
| | - | 偽陰性 (False negative) | 真陰性 (True negative) |

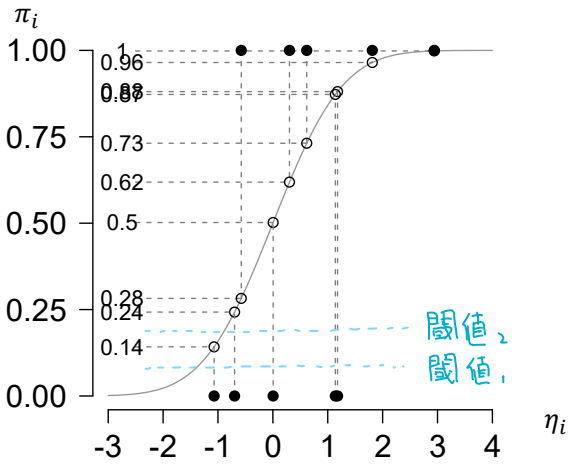
### ⑤ F1 スコア (F1 Score)

感度と精度の調和平均。

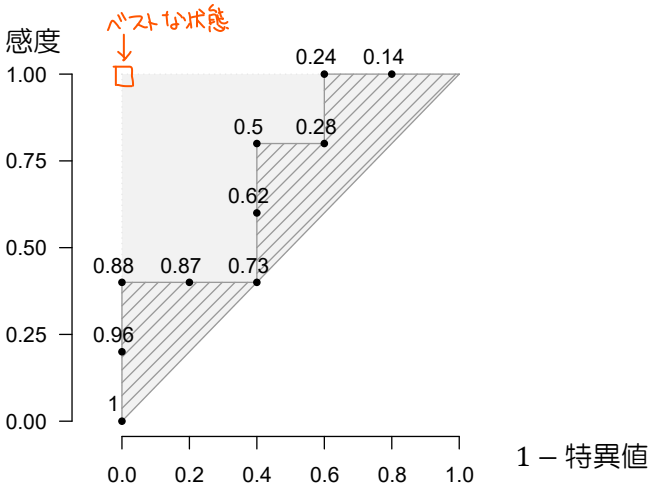
$$F_1 = \frac{2\,感度 \times 精度}{感度 + 精度}$$

## （4）ROC 曲線と AUC

① **ROC 曲線** Receiver Operative Characteristic Curve
それぞれの閾値に対して、特異値（横軸）、感度（縦軸）がどのくらいになるのかをプロットしたもの。




| 閾値 | 感度 | 1-特異度 | A | B | C | D |
|---|---|---|---|---|---|---|
| 0.000 | 1.0 | 1.0 | 5 | 0 | 5 | 0 |
| 0.142 | 1.0 | 0.8 | 5 | 0 | 4 | 1 |
| 0.243 | 1.0 | 0.6 | 5 | 0 | 3 | 2 |
| 0.282 | 0.8 | 0.6 | 4 | 1 | 3 | 2 |
| 0.502 | 0.8 | 0.4 | 4 | 1 | 2 | 3 |
| 0.619 | 0.6 | 0.4 | 3 | 2 | 2 | 3 |
| 0.731 | 0.4 | 0.4 | 2 | 3 | 2 | 3 |
| 0.873 | 0.4 | 0.2 | 2 | 3 | 1 | 4 |
| 0.881 | 0.4 | 0.0 | 2 | 3 | 0 | 5 |
| 0.965 | 0.2 | 0.0 | 1 | 4 | 0 | 5 |
| 0.998 | 0.0 | 0.0 | 0 | 5 | 0 | 5 |
| 1.000 | 0.0 | 0.0 | 0 | 5 | 0 | 5 |




Ⓟ 分類表

| | | 真の $y$ | |
|---|---|---|---|
| | | + | − |
| $\hat{y}$ | + | A | C |
| | − | B | D |

Ⓟ 閾値の値と分類表

(ケース1) 閾値1

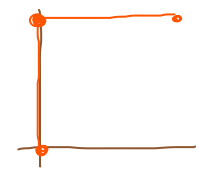
| | | 真の $y$ | |
|---|---|---|---|
| | | + | − |
| $\hat{y}$ | + | 5 | 5 |
| | − | 0 | 0 |

(ケース2) 閾値2

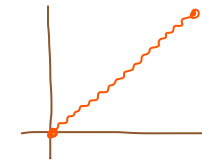
| | | 真の $y$ | |
|---|---|---|---|
| | | + | − |
| $\hat{y}$ | + | 5 | 4 |
| | − | 0 | 1 |

Ⓟ ROC 曲線のよしあし

(良い状態)

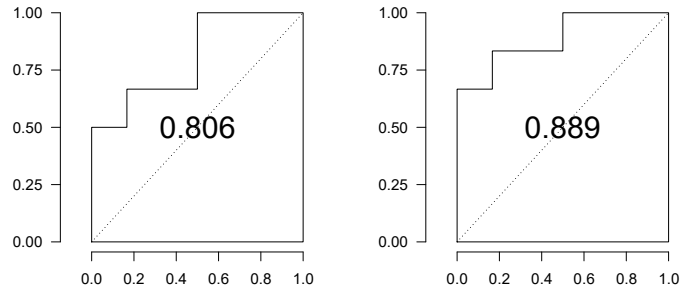


(悪い状態)



③ AUC　Area Under Curve

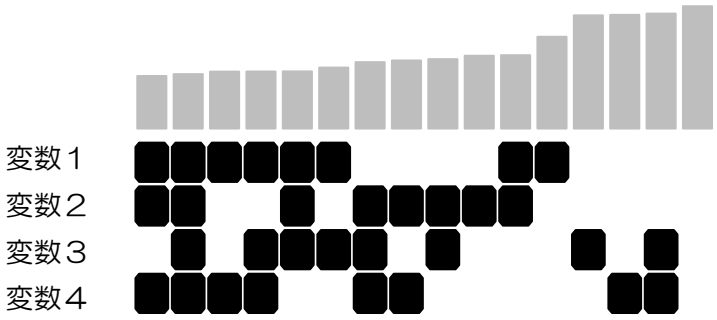
これは、ROC 曲線の下の面積で、複数のモデルの ROC 曲線のパフォーマンスを比較するために計算される。




⇒ クロスバリデーションと組みあわせて利用!

■ 情報量基準に基づくもの

重回帰同様、各種「情報量基準」を用いてモデル選択をする。変数の数が多ければ、ステップワイズ法などを検討する。




### モデル選択のポイント

① たくさんの**独立変数**を用意する

入れた独立変数の偏回帰係数しか推定できないので、研究の始めに、たくさんの候補を用意しておこう。

② **リンク関数**を変えてみる

モデルのパフォーマンスは、リンク関数の設定によっても変わる。いろいろ試してベストなものを探そう。

③ 母集団に想定する**モデルを変えてみる**

過分散やゼロ過剰を持つデータには統計モデルを変更して対応しよう。

④ **異なる基準**でモデル比較を行う

情報量基準もクロスバリデーションもあくまで一つの指標。複数検討して多角的に判断を下そう。

## ○ 考えたいこと 2：残差分析

重回帰では、予測値 $\hat{y}_i$ と観測値 $y_i$ の差である残差から特異なふるまいをするデータを検出した。

⇒ ロジスティック回帰でも、特異な点の有無を確認し、モデルの想定の妥当性を吟味しよう！

### （1）ピアソン残差 Pearson Residuals

$$X_k = \frac{y_k - n_k\hat{\pi}_k}{\sqrt{n_k\hat{\pi}_k(1-\hat{\pi}_k)}}$$

…期待値のまわりに

…標準偏差

何個分のところにいるのか

Ⓟ 二項分布 $Binom(n, \pi)$

$E[y] = n\pi$

$Var[y] = n\pi(1-\pi)$

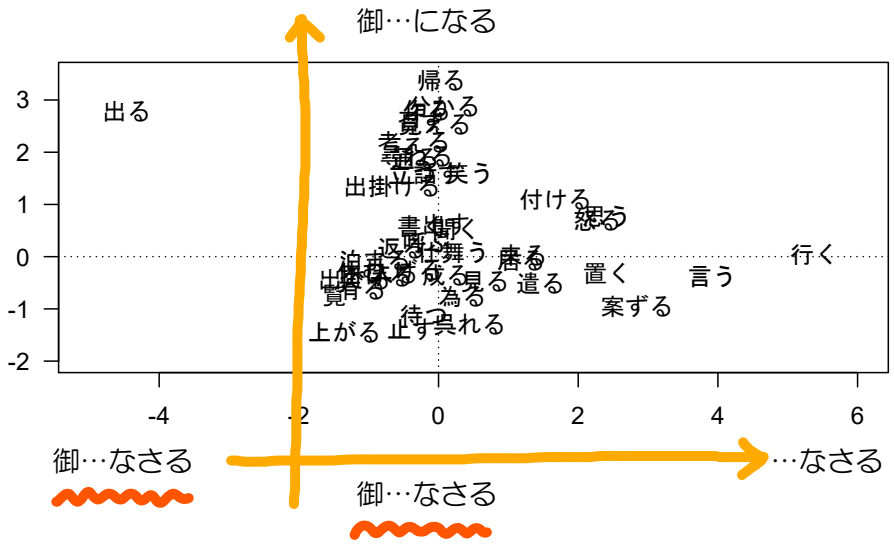
$Sd[y] = \sqrt{n\pi(1-\pi)}$

#### 標準化ピアソン残差 Standard Pearson Residuals

次のように標準化をして用いることも多い（$h_k$ はてこ値）。

$$r_{P_k} = \frac{X_k}{\sqrt{1-h_k}}$$

#### 例：江戸から昭和期の尊敬語の分析におけるピアソン残差



（2）逸脱度残差 Deviance Residuals

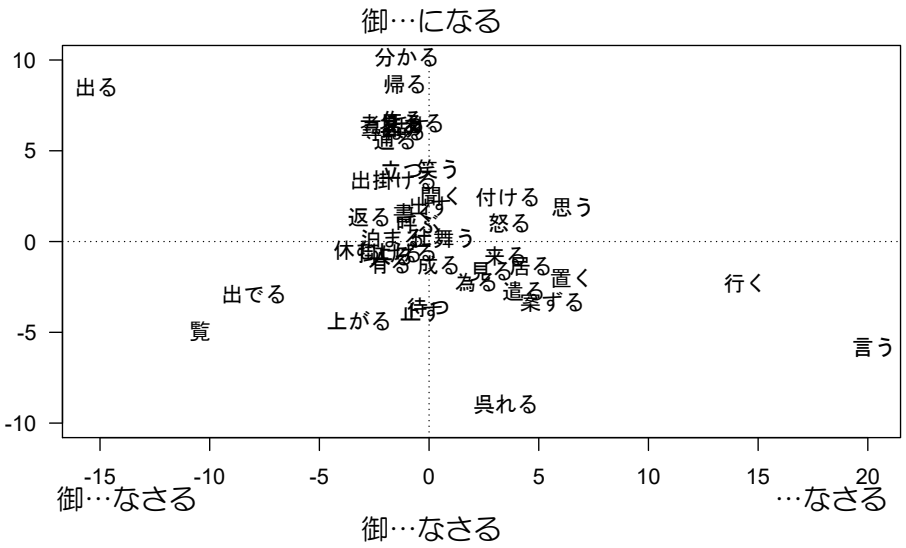
$$d_k = \text{sign}\,(y_k - n_k\hat{\pi}_k)\left\{2\left[y_k \log\left(\frac{y_k}{n_k\hat{\pi}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_k}\right)\right]\right\}^{1/2}$$

### 標準化逸脱度残差 Standard Deviance Residuals

次のように標準化をして用いることも多い（$h_k$はてこ値）。

$$r_{D_k} = \frac{d_k}{\sqrt{1 - h_k}}$$

### 例：江戸から昭和期の尊敬語の分析における逸脱度残差




| ID | $y_{i1}$ (御…になる) | $y_{i2}$ (…なさる) | $y_{i3}$ (御…なさる) | $x_{1i}$ 1 モーラ | $x_{2i}$ 意図性 |
|---|---|---|---|---|---|
| 1 分かる | 56 | 0 | 0 | 0 | 0 |
| 6 出る | 31 | 12 | 31 | 1 | 1 |
| 26 行く | 1 | 48 | 1 | 0 | 1 |
| 34 言う | 1 | 141 | 41 | 0 | 1 |
| 39 呉れる | 0 | 64 | 144 | 0 | 1 |
| 44 覧 | 90 | 0 | 319 | 0 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 合計 | 961 | 924 | 1599 | | |