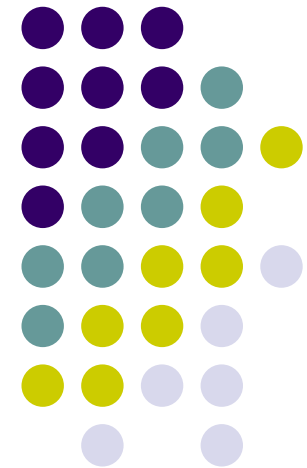
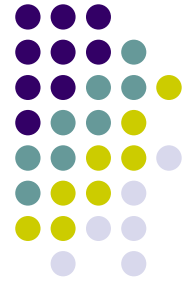


ロジスティック回帰分析 ～ 理論編

行動データ科学研究分野
M1 兼清 道雄



ロジスティック回帰分析 (logistic regression)とは



- 二値変数 (ex.)満足・不満足)
に対する回帰分析である



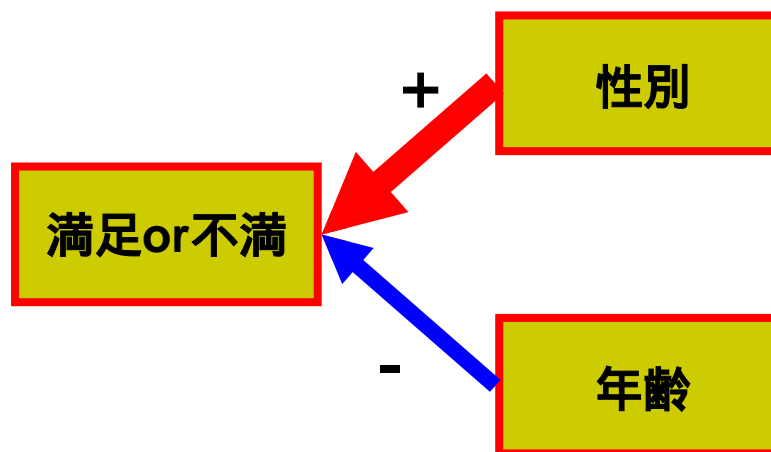
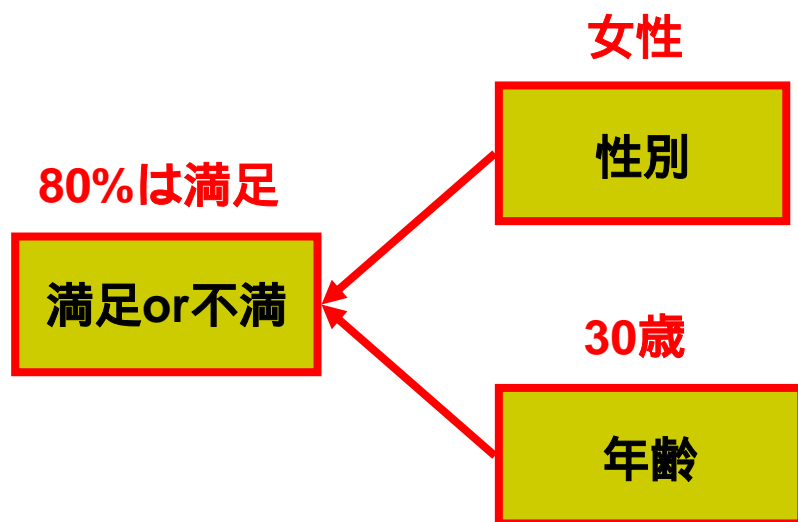
二値変数とは？

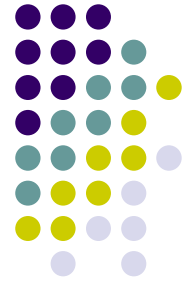
- 二つのカテゴリーからなるカテゴリカル変数
 - 消費税増税(5% 8%)に「賛成・反対」
 - ロジスティック回帰分析の使用経験「ある・ない」
 - 笑い飯が「好き・嫌い」
 - 笑い飯:M-1 グランプリ 2002,2003 決勝進出
 - 笑い飯を「知っている・知らない」
 - 逆上がりの「成功・失敗」
 - 行政に対して「満足・不満足」
- 0と1のデータと考えることが多い



ロジスティック回帰分析で出来る事

- 「性別」と「年齢」が与えられた時の「満足or不満」の予測
- 「性別」や「年齢」の影響の検討





普通の回帰分析ではダメ？

- 離散変数である二値変数を連続変数として予測するのは無理がある
 - データは「満足(y=1) or 不満(y=0)」
 - 予測式は

ID	満足	満足	性別	年齢
1	満足	1	女性	20
2	不満	0	男性	21
⋮				
99	満足	1	男性	37
100	満足	1	女性	38

$$\hat{y} = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

- xに値を代入した場合, 予測値y^は0 or 1 とは限らない
 - y^ = 0.8, -1.5, -0.2 ……??

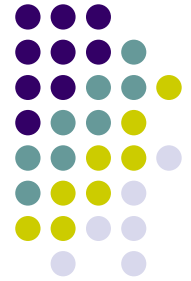


では確率で考えてみよう！

- $P(Y=1)$ に対して回帰モデルを当てはめる
 - 離散変数ではなく連続変数となる

$$P(Y = 1) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

- 線形確率モデル(linear probability model)
- でもダメ
- なぜ？
- 確率は[0,1]の範囲しか取らない
 - $p^{(Y=1)} = 1.5, -0.5$ などはどう解釈すればよい？
- じゃあ、どうすればいいだろう…



確率の「ロジット」を取ってみる！！

- ロジット(オッズの対数を取ったもの)

$$\text{logit}[P(Y = 1)] = \log \left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right]$$

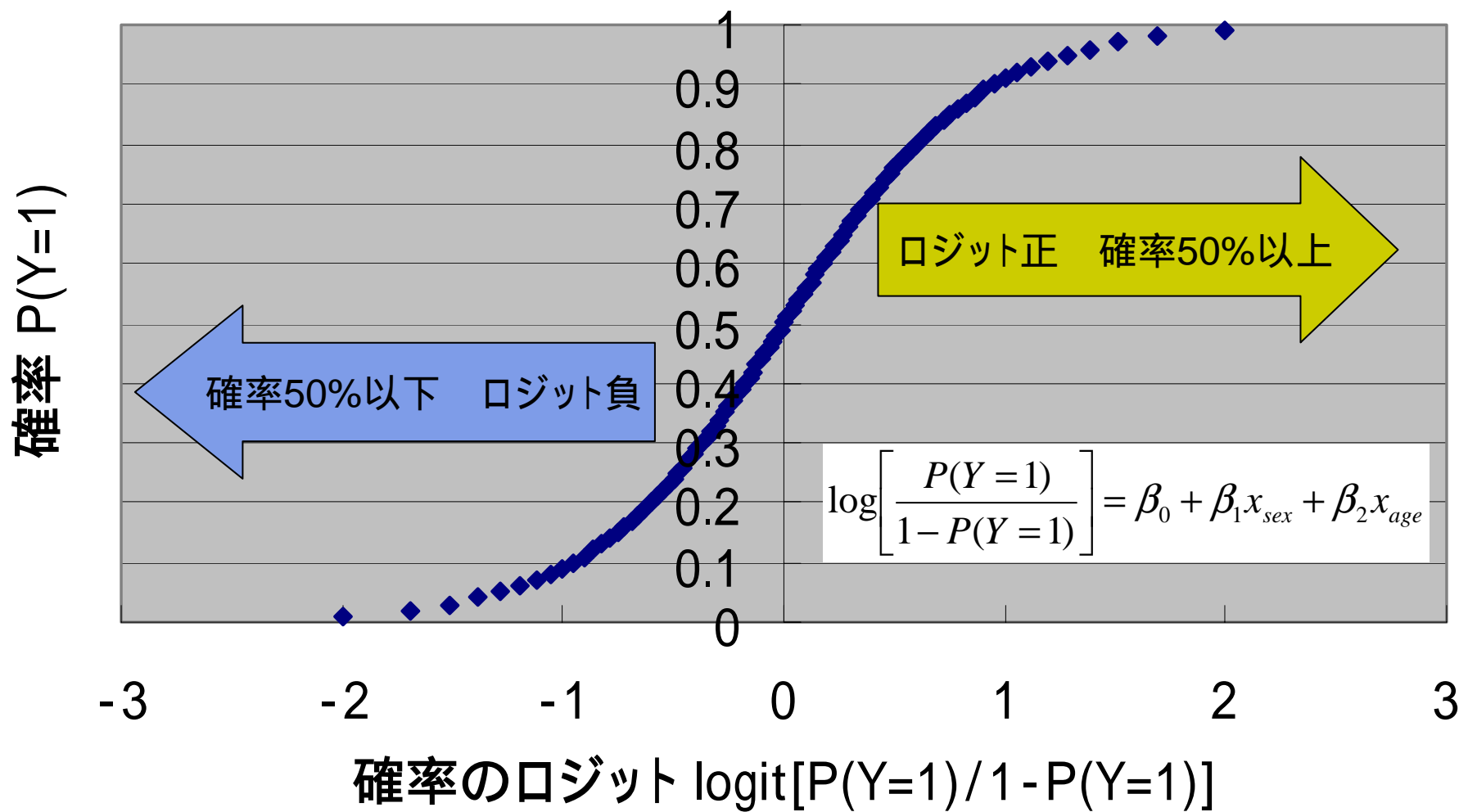
- こうすると、従属変数の範囲が[- , +]となる
- これに対して回帰モデルを当てはめる

$$\log \left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

- ロジットモデル(logit model)
= ロジスティック回帰モデル(logistic regression model) ₇



確率とそのロジットの関係





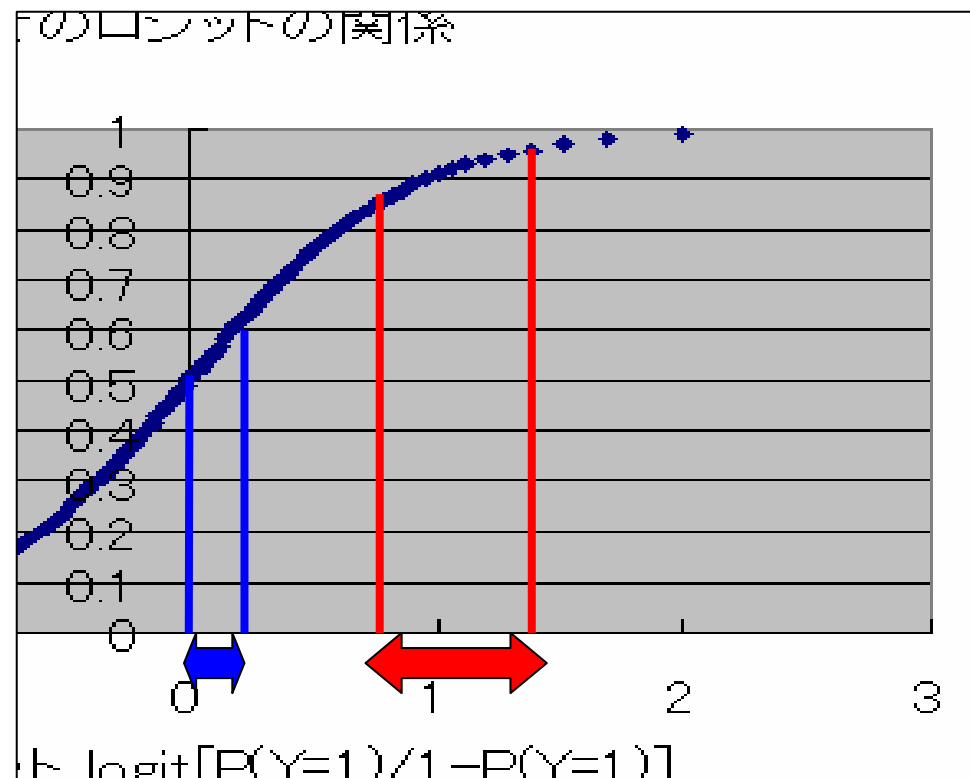
ロジットを用いる利点

- 範囲が[- , +]となる
- 確率に対する重み付けが実質科学的
- 偏回帰係数の解釈がしやすい

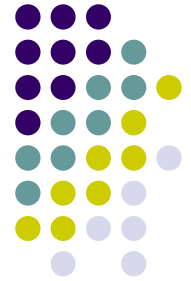


確率に対する重みづけ？

- 例えば満足確率を50% 60%に上げる場合と85% 95%と上げる場合では大変さは異なる
 - 治癒確率
 - 成功確率
- ロジットではそれが考慮される→
 - **50%** **60%**
 - **85%** **95%**



「頭がいい人の習慣術」 (小泉 2003)より引用



- 「なぜ、仕事で100点を目指そうとするのか」という節での一説

その理由は、学生時代のことを思い出してもらえば、ただちに了解されるはずだ。高校三年の初めに模擬試験で60点だった人が本番で80点とるということはザラにある話だが、さらに一年間浪人して80点を100点にするなどということは、まずありえない。

一年間必死に勉強しても、上乗せできるのは、せいぜい10点というところ。同じプラス20点でも、60点を80点にするための努力と、80点を100点にするための努力をくらべれば、後者のほうが数倍、努力が必要なのである。だから、浪人して急に実力が伸びるのは、高校時代に遊びすぎて、60点しかとれなかったような受験生と相場が決まっている。



ロジットを用いる利点

- 範囲が[- , +]となる
- 確率に対する重み付けが実質科学的
- 偏回帰係数の解釈がしやすい

ロジスティック回帰における 偏回帰係数の解釈について



- 以下の式の β_1 や β_2 は何を表すのか？

$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

- 両辺の指数をとると

$$\frac{P(Y=1)}{1-P(Y=1)} = \exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]$$

Pの1-Pに対するオッズ



1歳年上と「オッズ」で比較

- 「1歳年上(t+1)のオッズ」/「t歳のオッズ」

- オッズ比という

$$\frac{P(Y=1)}{1-P(Y=1)} = \exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]$$

$$\frac{\exp A}{\exp B} = \exp[A - B]$$

$$\frac{\frac{P(Y=1 | x_{age} = t+1)}{1 - P(Y=1 | x_{age} = t+1)}}{\frac{P(Y=1 | x_{age} = t)}{1 - P(Y=1 | x_{age} = t)}} = \frac{\exp[\beta_0 + \beta_1 x_{sex} + \beta_2 (t+1)]}{\exp[\beta_0 + \beta_1 x_{sex} + \beta_2 t]} = \exp\{\beta_2 (t+1) - \beta_2 t\} = \underline{\exp(\beta_2)}$$

- 偏回帰係数の指数を取ったもの = オッズ比



偏回帰係数の解釈

- 普通の回帰分析の偏回帰係数
 - 他の変数を条件づけたもとで, 当該変数が1増えたときのyの変化の差を表す
- ロジスティック回帰分析の偏回帰係数
 - 他の変数を条件づけたもとで, 当該変数が1増えたときの確率の変化の比を表す
 - オッズ比！！

オッズ比って？ そしてなぜオッズ比？





新薬の開発では・・・

- B薬(新薬)がA薬(従来薬)に比べて疾病の発症率を下げるものか否か?ということに興味あり

- p_0 と p_1 の比較に興味がある

	A薬	B薬
発症	p_0	p_1
未発症	$1-p_0$	$1-p_1$

- リスク差: $p_1 - p_0$
 - $0.15 - 0.20 = -0.05$

- リスク比: p_1 / p_0
 - $0.15 / 0.20 = 0.75$ (倍)

- オッズ比: $\{p_1 / (1 - p_1)\} / \{p_0 / (1 - p_0)\}$
 - $\{0.15 / (1 - 0.15)\} / \{0.20 / (1 - 0.20)\} = 0.71$ (倍)

	A薬	B薬
発症	20%	15%
未発症	80%	85%



リスク差・リスク比・オッズ比

- それぞれ特徴がある
- リスク差
 - どの%も同じ重み
 - 発症率10% 5%も, 90% 85%も同じリスク差で-5
- リスク比
 - 小さい%に重み大
 - 発症率10% 5%だと $5/10=0.5$ (半分に減った)
 - 発症率90% 85%だと $85/90=0.94..$ (ほとんど減らない)
- オッズ比
 - 極端な値(0%や100%)に対して重み大, 真ん中には重み小
 - 発症率10% 5%(オッズ比 0.47: 半分以下に減った)
 - 発症率90% 85%(オッズ比 0.63: まあまあ減った)

ロジスティック回帰の特徴, また,
「なぜロジスティックがいいのか?」,
について述べてきました
では, より詳しく見ていきましょう





目次

- パラメータ(偏回帰係数)の推定・検定
 - 効果はどれくらいか？効果はあるのか？
- 確率の推定
- モデルの適合
- モデルの比較
- エトセトラ

$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

パラメータの推定・検定

- パラメータの推定には最尤法が用いられる

- 最尤推定値は正規分布に従う

- ex.) $\hat{\beta}_2 \sim N(\beta_2, \text{Var}(\hat{\beta}_2))$

- パラメータの区間推定

- 信頼率95%の信頼区間

- ex.) $\hat{\beta}_2 \pm (1.96 \times \sqrt{\text{Var}(\hat{\beta}_2)})$

- $\hat{\beta}_2 = 0.50$, $\text{Var}(\hat{\beta}_2) = 0.10$ だと, $(0.304, 0.696)$

- オッズ比の区間推定は

- $(\exp(0.304), \exp(0.696))$ $(1.36, 2.01)$



- オッズ比の推定値 1.65
- オッズ比の区間推定 (1.36, 2.01)
 - 年齢が1歳上がるごとに1.65倍満足する
 - 年齢が1歳上がるごとに1.36倍 ~ 2.01倍満足する

$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

パラメータの推定・検定

● ワールド検定

- 推定値を標準誤差で割ったものの絶対値が1.96より大きければ5%水準で有意

- 推定値が正規分布に従うことから

- ex.)
$$z = \frac{\hat{\beta}_2}{\sqrt{\text{Var}(\hat{\beta}_2)}}$$

- $\hat{\beta}_2 = 0.50$, $\text{Var}(\hat{\beta}_2) = 0.10$ だと $z = 5 > 1.96$, よって有意

- 年齢による効果が有意にある！！

- また, z を2乗したものは自由度1の χ^2 分布に従う

- 2乗したものが3.84より大きければ有意

もう一つのパラメータの 検定方法

$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

尤度L1

$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex}$$

尤度L0

● 尤度比検定

- β_2 も含めたモデル(M1)の尤度L1と
 $\beta_2=0$ としたモデル(M0)の尤度L0を比較する
- $-2\log(L0/L1)=-2[\log(L0)-\log(L1)]$ が(帰無仮説の下
で)自由度1の χ^2 分布に従う
 - 帰無仮説はM0=M1 ($\beta_2=0$)
 - ex.) $\log(L1)=220$, $\log(L0)=217$
 $-2[\log(L0)-\log(L1)]=6.0>3.84$ より有意!
 - $\beta_2 = 0$ 年齢の効果あり!!
- ワルド検定よりも検出力が高い!



目次

- パラメータ(偏回帰係数)の推定・検定
- 確率の推定
 - ある条件における確率は？
- モデルの適合
- モデルの比較
- エトセトラ

$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

確率の推定

- 例えば, 男性で30歳の人の満足と答える確率を知りたい!! ということもある
- 前述の式

$$\frac{P(Y=1)}{1-P(Y=1)} = \exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]$$

- 変形すると

$$\begin{aligned} P(Y=1) &= \frac{\exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]}{1 + \exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]} \\ &= \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age})]} \end{aligned}$$

推定値を代入
&
説明変数を代入



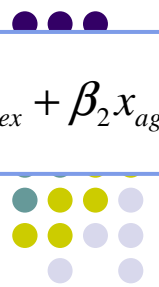
確率の推定

- ex.) $\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$ にそれぞれ代入
 $\beta_0 + \beta_1 \times 0 + \beta_2 \times 30 = 1.5$

$$P(Y=1) = \frac{\exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]}{1 + \exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]}$$

代入すると

- $P(Y=1) = 0.82$
 - 男性30歳の満足する確率は82% !


$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

確率の区間推定

- 推定されたロジットのバラツキ

$$\begin{aligned} \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{sex} + \hat{\beta}_2 x_{age}) &= \text{Var}(\hat{\beta}_0) + x_{sex}^2 \text{Var}(\hat{\beta}_1) + x_{age}^2 \text{Var}(\hat{\beta}_2) \\ &\quad + 2x_{sex} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + 2x_{age} \text{Cov}(\hat{\beta}_0, \hat{\beta}_2) + 2x_{sex}x_{age} \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \end{aligned}$$

- ロジットの信頼率95%の信頼区間

$$(\hat{\beta}_0 + \hat{\beta}_1 x_{sex} + \hat{\beta}_2 x_{age}) \pm 1.96 \times \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{sex} + \hat{\beta}_2 x_{age})}$$

- ex.) 男性30歳のロジットの区間推定 (1.20, 2.00)

前回から修正あり！！



確率の区間推定

- 続き...

- ex.) 男性30歳のロジットの区間推定 (1.20,2.00)
 - それぞれの値(1.20と2.00)を確率を推定する際に使用した式に代入

$$P(Y = 1) = \frac{\exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]}{1 + \exp[\beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}]}$$

ここに代入

- 男性30歳が満足する確率の区間推定(0.77,0.88)



目次

- パラメータ(偏回帰係数)の推定・検定
- 確率の推定
- モデルの適合
 - そもそもモデルはあっている？
 - 間違っていたらパラメータも,
それを基に算出した確率も意味を持たない
- モデルの比較
- エトセトラ

$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_2 x_{age}$$

適合度検定

- 簡単にするため年齢のみのモデルを考える
- モデルから得られる予測確率より当てはめ度数を得る

- 25-29歳
 - 25歳: 10人 × 予測確率30% = 3人
 - 27歳: 10人 × 予測確率40% = 4人
 - 29歳: 10人 × 予測確率50% = 5人

- よって当てはめ度数は12人

- 以下の2つの値

$$X^2 = \sum \frac{(\text{観測度数} - \text{当てはめ度数})^2}{\text{当てはめ度数}}$$

$$G^2 = 2 \sum (\text{観測度数}) \log \left(\frac{\text{観測度数}}{\text{当てはめ度数}} \right)$$

	満足 観測度数	満足当て はめ度数	カテゴリの 中の度数
-19歳	6	5	20
20-24歳	8	9	25
25-29歳	10	12	30
30-34歳	17	15	28
35-39歳	19	17	27
40歳-	16	20	29

- これらが、カテゴリ (6つ) - パラメータ (2つ) = 4の自由度を持つ分布に従うことを利用する
- 大きく非有意であれば、適合していると考えられる
 - ex.) $X^2 = 1.95$ (p=.745) よって、よく適合していると考えられる



適合度検定

- 連続変数をグルーピングする理由
 - セルの当てはめ度数が5より小さい場合, それぞれの値 X^2 , G^2 は χ^2 分布に従わなくなるため
- しながら説明変数が多い場合はセルが「ごっつ」多い分割表となってしまう
 - 年齢・性別・収入・家族構成・云々
 - ほとんどのセルの度数が小さくなる
 - 必然的に前述検定が出来なくなる
- Hosmer-Lemeshow検定が代替案
 - 予測確率順でグルーピングする方法
 - Hosmer and Lemeshow (1989, p140)参照

予測確率	観測	予測	
1 ~ 10位			
11 ~ 20位			
⋮			
91 ~ 100位			

Hosmer-Lemeshow test



その他にも

- 尤度比検定(後述)
- ピアソン残差(カテゴリーカル:p158-)
- 影響診断(カテゴリーカル:p160-)
- 感度分析
 - 特に外挿の問題に対して
- で, モデルが適合しているかどうかを見ることが出来る



目次

- パラメータ(偏回帰係数)の推定・検定
- 確率の推定
- モデルの適合
- モデルの比較
 - よりよいモデルを目指して・・・
- エトセトラ

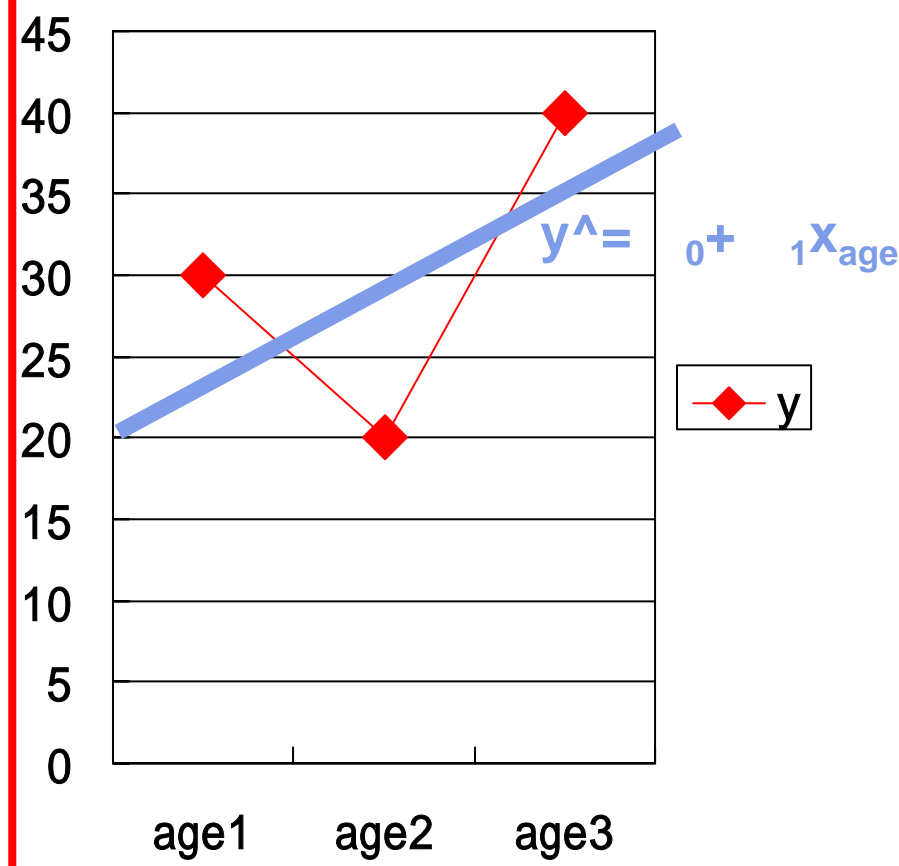
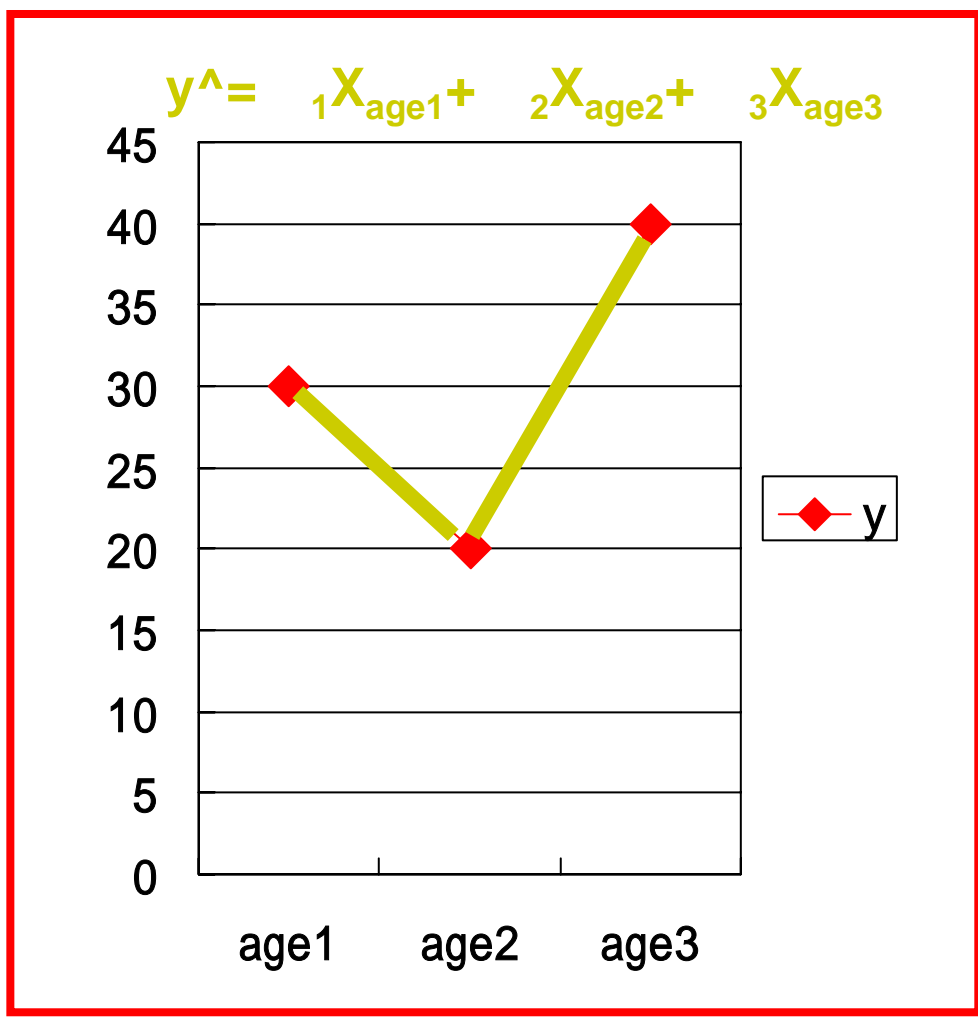
$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age}$$

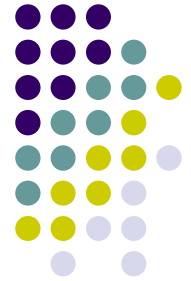
尤度比検定

- 飽和モデルの尤度 L_S と
当該モデル L_1 の尤度の比較
 - 飽和モデル: データに完全にフィットしたモデル
 - この尤度比検定統計量 D を「デビアンス」と呼ぶ
 - $D = -2\log[L_1/L_S] = -2[\log(L_1) - \log(L_S)]$
 - D は自由度が「飽和モデルのパラメータ数
- 当該モデルのパラメータ数」の χ^2 分布に従う
- 非有意ならば, 当該モデルが飽和モデルより
適していることになる



飽和モデルとは(ex.)回帰分析)





尤度比検定

- 同じ要領で・・・
- ex.) 年齢・性別・年齢 * 性別のモデルと性別のみのモデルを比べることも可能

$$\log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age} + \beta_3 x_{sex} x_{age} \quad \text{v.s.} \quad \log \left[\frac{P(Y=1)}{1-P(Y=1)} \right] = \beta_0 + \beta_1 x_{sex}$$

- 尤度比検定統計量：
-2{log(性別のみの尤度)-log(年・性・年 * 性の尤度)}
 - パラメータ数の差は2
 - 自由度2の χ^2 分布に従う
- 有意：性別のみに簡略化してはいけない
- 非有意：性別のみに簡略化出来る



よりよいモデル

- 飽和モデルくらい適合している
 - 飽和モデルとの尤度比検定が非有意
- パラメータは少ない方がよい
- 実質科学的観点からのモデル選択が必要
 - 意味無くパラメータを減らさない
 - 何故そのパラメータは必要ないのかという理由付け
 - 実質科学的に意味のあるパラメータの場合、残すという方法もある



目次

- パラメータ(偏回帰係数)の推定・検定
- 確率の推定
- モデルの適合
- モデルの比較
- エトセトラ



エトセトラ

- 今回取り上げられなかったが重要だと思うこと
 - プロファイル尤度による信頼区間
 - ワールド検定による区間推定よりも良い
 - 条件付き最尤法
 - 「データがアンバランス」などによる小標本に対して効果的
 - サンプルサイズと検出力
 - カテゴリカルデータ解析入門(p179)
 - 5.6ロジスティック回帰におけるサンプルサイズと検出力



参考文献・参考HP・参考資料

- カテゴリカルデータ解析入門
 - Alan Agresti 著 渡邊ら訳 (2003) サイエンティスト社
 - 第4章 一般化線形モデル 第5章 ロジスティック回帰
- ロジスティック回帰分析
 - 丹後俊郎・山岡和枝・高木晴良 著 (1996) 朝倉書店
- ロジスティック回帰分析の発表 by 鳥居氏
 - <http://koko15.hus.osaka-u.ac.jp/~torii/donut.html>
 - 狩野先生の資料もここから手に入る
- ロジスティック回帰入門
 - <http://www.h5.dion.ne.jp/~ge3j-ari/stat/logis.html>
- ロジスティック回帰分析(上記 + の発表スライド)
 - <http://kyoumu.educ.kyoto-u.ac.jp/cogpsy/personal/Kusumi/datasem03/hirayama.files/frame.htm>
- 山本氏(行動計量学B3)の発表資料
- エトセトラ...