OpenIntro

第 9 章：重回帰とロジスティック回帰

OpenIntro Statistics 第 4 版（日本語版）

原著スライド：Mine Çetinkaya-Rundel（OpenIntro）

重回帰

- 単純線形回帰：2 変数 — 応答変数 y と説明変数 x
- 重線形回帰：複数変数 — 応答変数 y と説明変数 $x_1, x_2, \cdots$

貧困率 vs. 地域（東部・西部）

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- 説明変数：地域、参照水準：東部（east）
- 切片：東部の州における貧困率の推定平均は 11.17%
  - 説明変数に *0* を代入したときの値
- 傾き：西部の州における貧困率の推定平均は東部より 0.38% 高い。
  - したがって、西部の州の貧困率の推定平均は 11.17 + 0.38 = 11.55%。
  - 説明変数に *1* を代入したときの値

貧困率 vs. 地域（北東部・中西部・西部・南部）

北東部（northeast）・中西部（midwest）・西部（west）・南部（south）のうち、参照水準はどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast（北東部）
(b) midwest（中西部）
(c) west（西部）
(d) south（南部）
(e) 判別できない

貧困率 vs. 地域（北東部・中西部・西部・南部）

北東部・中西部・西部・南部のうち、最も貧困率が低い地域はどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast（北東部）
(b) midwest（中西部）
(c) west（西部）
(d) south（南部）
(e) 判別できない

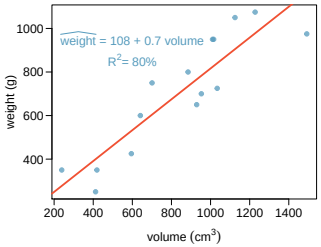
本の重量

	重量 (g)	体積 (cm^3)	カバー
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) “Data Analysis and Graphics Using R”

本の重量（続き）

散布図は本の重量と体積の関係、および回帰出力を示している。以下のうち正しいものはどれか？



(a) このモデルで本の 80% の重量を正確に予測できる。
(b) 体積が平均より 10 cm^3 大きい本は、重量が平均より約 7 g 重いと予測される。
(c) 重量と体積の相関係数は $R = 0.80^2 = 0.64$ である。
(d) このモデルは最も体積の大きい本の重量を過小推定している。

体積を用いた本の重量のモデリング

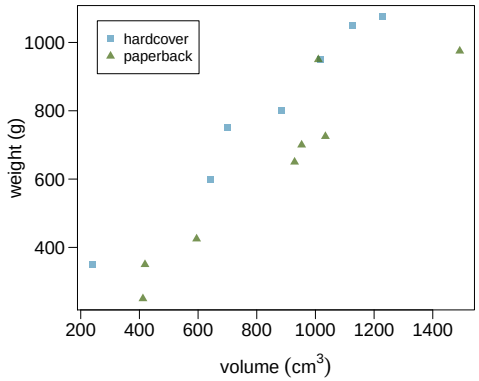
出力の一部を省略...

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06


Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared: 0.8026, Adjusted R-squared: 0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

ハードカバーとペーパーバックの重量

ハードカバーとペーパーバックの体積と重量の関係において、何かトレンドを見つけられるか？



体積とカバー種類を用いた本の重量のモデリング

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***


Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275, Adjusted R-squared: 0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```


参照水準の決定

以下の回帰出力に基づき、変数 `cover` の参照水準はどれか？ なお `pb`：ペーパーバック。

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) ペーパーバック（paperback）
(b) ハードカバー（hardcover）

参照水準の決定

以下のうち、この回帰モデルにおける変数の役割を正しく説明しているものはどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) 応答変数：重量、説明変数：体積、ペーパーバック
(b) 応答変数：重量、説明変数：体積、ハードカバー
(c) 応答変数：体積、説明変数：重量、カバー種類
(d) 応答変数：重量、説明変数：体積、カバー種類

線形モデル

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

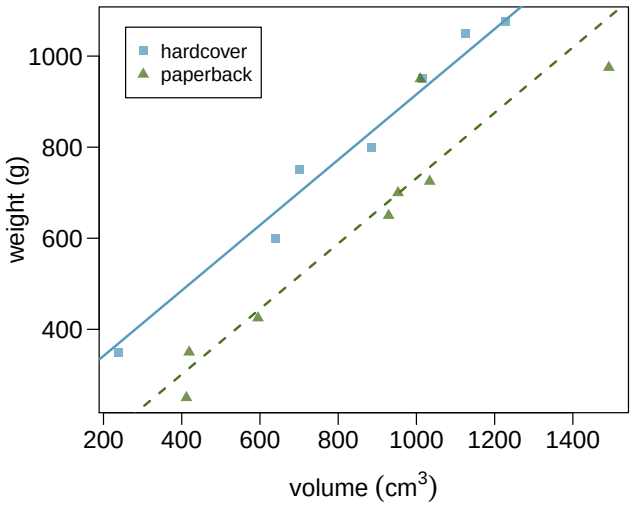
1. ハードカバーの場合：`cover` に *0* を代入

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

2. ペーパーバックの場合：`cover` に *1* を代入

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 1 \\ &= 13.91 + 0.72\ volume \end{aligned}$$

線形モデルの可視化



回帰係数の解釈

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- 体積の傾き：他の変数を一定に保つと、体積が 1 cm^3 大きい本は重量が約 0.72 g 重い傾向がある。
- カバーの傾き：他の変数を一定に保つと、ペーパーバックはハードカバーより 184 g 軽いとモデルは予測する。
- 切片：体積ゼロのハードカバーの平均重量は 198 g と予測される。
  - 明らかに、この切片は文脈的に意味をなさない。直線の高さを調整するためだけのものである。

予測

体積が 600 cm^3 のペーパーバックの予測重量を求める正しい計算式はどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) 197.96 + 0.72 * 600 - 184.05 * 1 $= 445.91$ グラム
(b) 184.05 + 0.72 * 600 - 197.96 * 1
(c) 197.96 + 0.72 * 600 - 184.05 * 0
(d) 197.96 + 0.72 * 1 - 184.05 * 600

別の例：子供の認知テストスコアのモデリング

3〜4 歳の子供の認知テストスコアを母親の特徴で予測する。データは米国成人女性とその子供を対象とした調査（全国縦断的青年調査のサブサンプル）。

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* (2007) Cambridge University Press.

傾きの解釈

母親の IQ の傾きの正しい解釈はどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

傾きの解釈

切片の正しい解釈はどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

傾きの解釈

`mom_work` の傾きの正しい解釈はどれか？

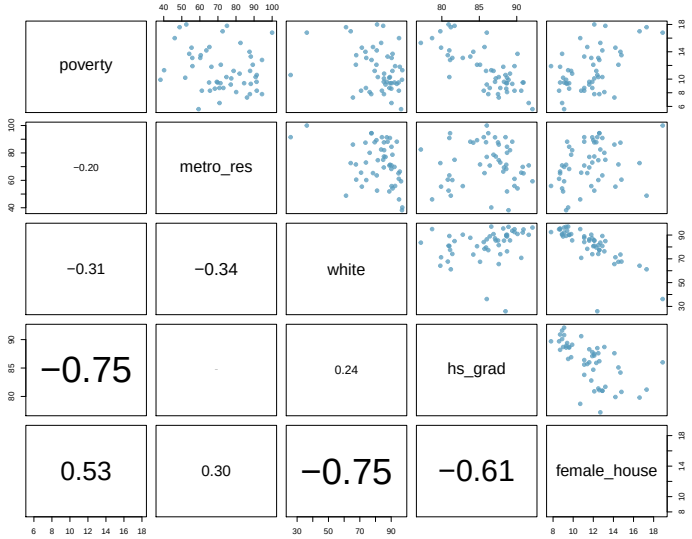
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

他の条件が等しい場合、子供の生後 3 年間に母親が働いていた子供は、働いていなかった子供と比べて

(a) スコアが 2.54 点低いと推定される

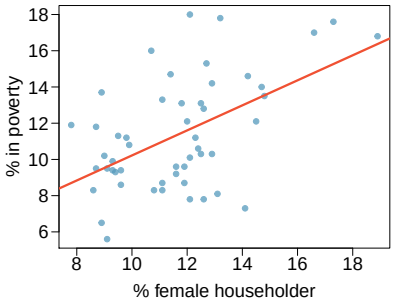
(b) スコアが 2.54 点高いと推定される

復習：貧困率のモデリング



% 女性世帯主を用いた貧困率の予測

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00



$$R = 0.53$$
$$R^2 = 0.53^2 = 0.28$$

R^2 を改めて見る

R^2 は 3 通りの方法で計算できる：

1. x と y の相関係数を 2 乗する（これまでの計算方法）
2. y と $\hat{y}$ の相関係数を 2 乗する
3. 定義に基づいて計算する：

$$R^2 = \frac{y \text{ の説明された変動}}{y \text{ の全変動}}$$

分散分析（*ANOVA*）を用いると、y の説明された変動と全変動を計算できる。

平方和

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$\begin{aligned} y \text{ の平方和：} SS_{Total} &= \sum (y - \bar{y})^2 = 480.25 \rightarrow \text{全変動} \\ \text{残差平方和：} SS_{Error} &= \sum e_i^2 = 347.68 \rightarrow \text{説明されていない変動} \\ x \text{ の平方和：} SS_{Model} &= SS_{Total} - SS_{Error} \rightarrow \text{説明された変動} \\ &= 480.25 - 347.68 = 132.57 \end{aligned}$$

$$R^2 = \frac{\text{説明された変動}}{\text{全変動}} = \frac{132.57}{480.25} = 0.28 \checkmark$$

なぜ別の方法が必要か？

相関係数の 2 乗として R^2 を計算する完全な方法があるのに、なぜ別のアプローチが必要なのか？

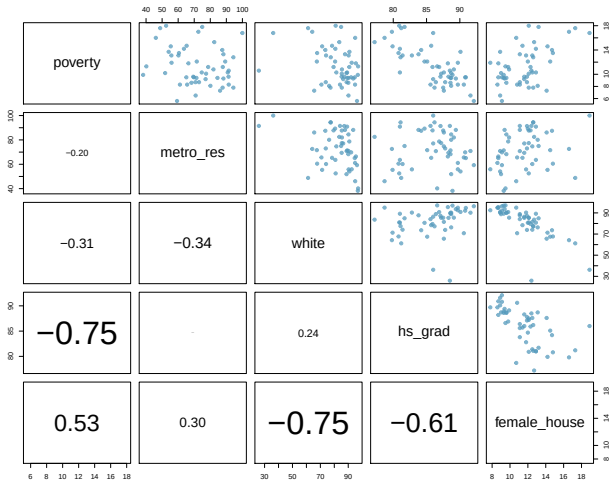
% 女性世帯主 + % 白人を用いた貧困率の予測

線形モデル：	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA：	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{\text{説明された変動}}{\text{全変動}} = \frac{132.57 + 8.21}{480.25} = 0.29$$

変数 `white` をモデルに追加することで、`female_house` が提供していなかった有益な情報が加わるか？



説明変数間の多重共線性

貧困率 *vs.* % 女性世帯主

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	*0.69*	0.16	4.32	0.00

貧困率 *vs.* % 女性世帯主 と % 白人

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	*0.89*	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

説明変数間の多重共線性（続き）

- 2つの予測変数が相関しているとき、それらは共線性があると言い、この多重共線性はモデルの推定を複雑にする。
  復習：予測変数は説明変数あるいは独立変数とも呼ばれる。理想的には互いに独立であるべきである。
- 互いに関連した予測変数をモデルに加えることは好ましくない。そのような変数を追加しても、多くの場合何も得られないからである。代わりに、最もシンプルで優れたモデル、すなわち簡潔な（節約的な）モデルを目指す。
- 観測データから生じる多重共線性を完全に避けることは不可能だが、実験は通常、予測変数間の相関を防ぐように設計される。

R^2 vs. 自由度調整済み R^2

	R^2	自由度調整済み R^2
モデル 1（単一予測変数）	0.28	0.26
モデル 2（重回帰）	0.29	0.26

- どんな変数をモデルに加えても R^2 は増加する。
- しかし、加えた変数が実際に新しい情報を提供しない場合や全く無関係な場合、自由度調整済み R^2 は増加しない。

自由度調整済み R^2

自由度調整済み R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

ここで n はケース数、p はモデル内の予測変数（説明変数）の数。

- p が負になることはないため、R^2_{adj} は常に R^2 より小さい。
- R^2_{adj} はモデルに含まれる予測変数の数に対してペナルティを課す。
- したがって、R^2_{adj} が高いモデルをより好ましいモデルとして選ぶ。

自由度調整済み R^2 の計算

ANOVA：	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$\begin{aligned} R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\ &= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\ &= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\ &= 1 - 0.74 \\ &= 0.26 \end{aligned}$$

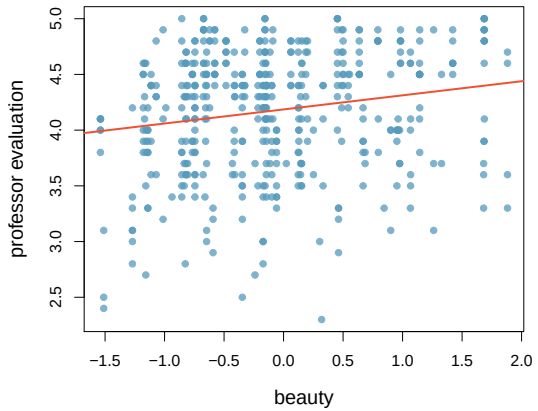
教室における美貌

- データ：テキサス大学の 463 コースにおける教員の美貌と授業評価の学生評価。
- 評価は学期末に実施され、美貌の判定はその後、授業に参加しておらず授業評価を知らない 6 人の学生（上位学年の女性 2 人、男性 2 人、下位学年の女性 1 人、男性 1 人）が行った。

Hamermesh & Parker. (2004)"Beauty in the classroom: instructors? pulchritude and putative pedagogical productivity? Economics Education Review.

教員評価 vs. 美貌

教員評価スコア（高いほど良い）vs. 美貌スコア（0 が平均、負が平均以下、正が平均以上）：



モデル出力に基づき、以下のうち正しいものはどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.19	0.03	167.24	0.00
beauty	0.13	0.03	4.00	0.00

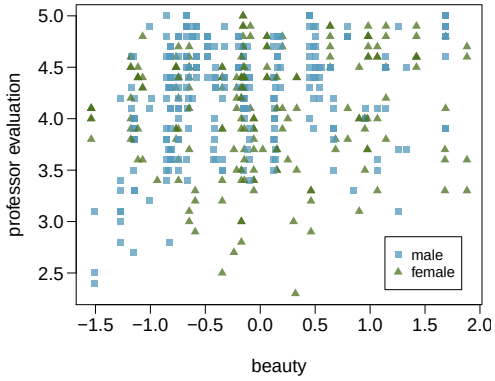
$R^2 = 0.0336$

(a) このモデルは教員評価の 3.36% を正確に予測できる。
(b) 美貌は教員評価の有意な予測変数ではない。
(c) 美貌スコアが平均より 1 点高い教員は、評価も 0.13 点高い傾向がある。
(d) 美貌スコアの変動の 3.36% は教員評価で説明できる。
(e) 相関係数は $\sqrt{0.0336} = 0.18$ または -0.18 の可能性があるが、どちらか判別できない。

探索的分析

興味深い特徴はあるか？

同じ美貌スコアを持つ場合、男性教員は女性教員より高く評価されるか、低く評価されるか、ほぼ同じか？



教員評価 vs. 美貌 + 性別

同じ美貌スコアを持つ場合、男性教員は女性教員より高く評価されるか、低く評価されるか、ほぼ同じか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00

$R^2_{adj} = 0.057$

(a) 高い → 美貌を一定にすると、男性教員は女性教員より平均 *0.17* 点高く評価される。

(b) 低い

(c) ほぼ同じ

フルモデル

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes [1]	0.1511	0.0749	2.02	0.04
lower.yes [2]	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students [3]	-0.0004	0.0004	-1.03	0.30
tenure.tenure track [4]	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

[1] `formal`：ネクタイ・ジャケット/ブラウスを着た写真、水準：`yes`、`no`
[2] `lower`：下位学年のコース、水準：`yes`、`no`
[3] `students`：学生数
[4] `tenure`：テニュア状況、水準：`non-tenure track`、`tenure track`、`tenured`

仮説

傾きパラメータの解釈がモデル内の他の変数を考慮に入れるように、予測変数の有意性をテストするための仮説も他の変数を考慮に入れる。

H_0 : 他の説明変数がモデルに含まれるとき、$B_i = 0$。

H_A : 他の説明変数がモデルに含まれるとき、$B_i \neq 0$。

有意性の評価：数値変数

年齢の p 値が 0.01 である。これは何を意味するか？

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) p 値が正なので、教員の年齢が高いほど評価も高いと予測される。

(b) 他の変数をモデルに保ったとき、教員の年齢が評価と関連しているという強い証拠がある。

(c) 年齢の真の傾きパラメータが 0 である確率は 0.01 である。

(d) 年齢の真の傾きパラメータが -0.0089 である確率が約 1% ある。

有意性の評価：カテゴリ変数

テニュアは 3 水準のカテゴリ変数（non tenure track、tenure track、tenured）である。以下のモデル出力に基づき、誤りはどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) 参照水準は non tenure track である。

(b) 他の条件が等しい場合、tenure track の教員は non-tenure track の教員より平均 0.19 点低く評価される。

(c) 他の条件が等しい場合、tenured の教員は non-tenure track の教員より平均 0.16 点低く評価される。

(d) 他の条件が等しい場合、tenure track と tenured の教員の平均評価には有意な差がある。

有意性の評価

教員の評価スコアの有意な予測変数ではない可能性があるなど、モデルへの貢献が少ないと思われる予測変数はどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes	0.1511	0.0749	2.02	0.04
lower.yes	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students	-0.0004	0.0004	-1.03	0.30
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

モデル選択戦略

これまでに学んだことを踏まえ、どの変数をモデルに保ち、どれを除くかを決めるために使える方法をいくつか考えてみよ。

後退消去法

1. フルモデルから始める
2. 1 つの変数を除いて各小モデルの R^2_{adj} を記録する
3. R^2_{adj} が最も大きく増加したモデルを選ぶ
4. いずれのモデルも R^2_{adj} が増加しなくなるまで繰り返す

後退消去法

フル	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
ステップ 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
ステップ 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	*0.0846*
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
ステップ 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
	beauty + gender + age + formal + native + tenure	*0.0855*
	beauty + gender + age + formal + native + students	0.0713
ステップ 4	gender + age + formal + native + tenure	0.0667
	beauty + age + formal + native + tenure	0.0553
	beauty + gender + formal + native + tenure	0.0723
	beauty + gender + age + native + tenure	0.0806
	beauty + gender + age + formal + tenure	0.0773
	beauty + gender + age + formal + native	0.0713

R の step 関数

R の step 関数は同様の後退消去プロセスを行うが、モデル選択に R^2_{adj} の代わりに AIC（赤池情報量基準）という異なる指標を使用する。

```
Call:
lm(formula = profevaluation ~ beauty + gender + age + formal +
    native + tenure, data = d)

Coefficients:
       (Intercept)              beauty          gendermale
          4.628435            0.105546            0.208079
                 age           formalyes   nativenon english
         -0.008844            0.132422           -0.243003
tenuretenure track       tenuretenured
         -0.206784           -0.175967
```

最良モデル：beauty + gender + age + formal + native + tenure

前進選択法

1. 応答変数 vs. 各説明変数の回帰から始める
2. R^2_{adj} が最も高いモデルを選ぶ
3. 残りの変数を 1 つずつ既存のモデルに追加し、R^2_{adj} が最も高いモデルを選ぶ
4. 残りの変数を追加しても R^2_{adj} が増加しなくなるまで繰り返す

- p 値アプローチによる後退消去法：
  1. フルモデルから始める
  2. p 値が最も高い変数を除いて小さいモデルを再適合する
  3. モデルに残っているすべての変数が有意になるまで繰り返す
- p 値アプローチによる前進選択法：
  1. 応答変数 vs. 各説明変数の回帰から始める
  2. 最も低い有意な p 値を持つ変数を選ぶ
  3. 残りの変数を 1 つずつ既存のモデルに追加し、最も低い有意な p 値を持つ変数を選ぶ
  4. 残りの変数が有意な p 値を持たなくなるまで繰り返す

後退消去法：p 値アプローチ

ステップ	含まれる変数と p 値									
フル	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
		male		yes	yes	nonenglish	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
ステップ 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
		male		yes	yes	nonenglish			tenure track	tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01
ステップ 2	beauty	gender	age	formal		native		students	tenure	tenure
		male		yes		nonenglish			tenure track	tenured
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.01
ステップ 3	beauty	gender	age	formal		native			tenure	tenure
		male		yes		nonenglish			tenure track	tenured
	0.00	0.00	0.01	*0.06*		0.02			0.01	0.01
ステップ 4	beauty	gender	age			native			tenure	tenure
		male				nonenglish			tenure track	tenured
	0.00	0.00	0.01			*0.06*			0.01	0.01
ステップ 5	beauty	gender	age						tenure	tenure
		male							tenure track	tenured
	0.00	0.00	0.01						0.01	0.01

最良モデル：beauty + gender + age + tenure

自由度調整済み R^2 vs. p 値アプローチ

- 2 つのアプローチは似ているが、異なるモデルになることがあり、自由度調整済み R^2 アプローチの方が最終モデルに含める予測変数が多くなる傾向がある。
- 予測精度の向上のみが目標であれば R^2 を使う。機械学習の応用でよく見られる。
- 応答変数の統計的に有意な予測変数がどれかを理解することに関心がある場合、または予測精度をわずかに犠牲にしても単純なモデルを作りたい場合は p 値アプローチが好まれる。
- どのアプローチを使っても、変数選択後の作業はまだ終わっていない — モデルの条件（仮定）が妥当かどうかを確認する必要がある。

モデルの条件（仮定）

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

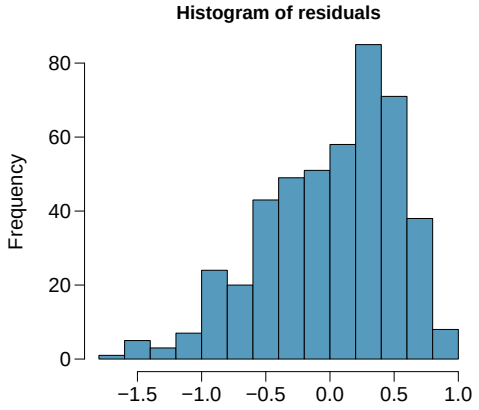
モデルは以下の条件に依存する

1. 残差がほぼ正規分布に従う（大きなデータセットでは重要性が低い）
2. 残差の分散が一定である
3. 残差が独立である
4. 各変数が応答変数と線形の関係にある

これらの条件の妥当性を確認するためにグラフによる方法をよく用いる。以下のスライドで詳しく解説する。

（1）ほぼ正規分布に従う残差

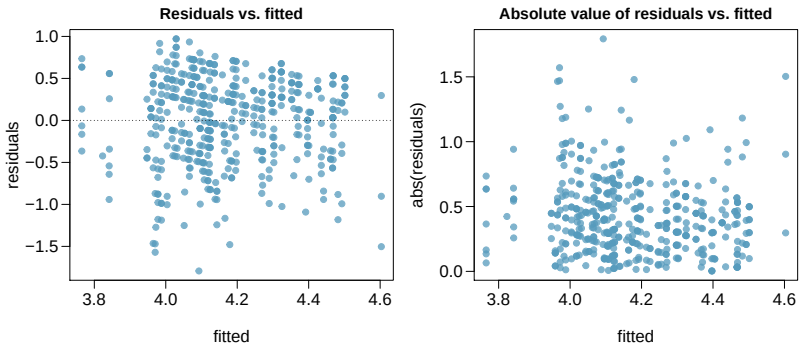
残差の正規確率プロットおよび/またはヒストグラム：



この条件は満たされているように見えるか？

(2) 残差の分散が一定

残差および/または残差の絶対値 vs. 当てはめ値（予測値）の散布図：



この条件は満たされているように見えるか？

一定分散の確認 — まとめ

- 単純線形回帰（1 つの説明変数）では、残差 *vs.* x のプロットを用いて一定分散の条件を確認した。
- 重線形回帰（2 つ以上の説明変数）では、残差 *vs.* 当てはめ値 のプロットを用いて一定分散の条件を確認する。

なぜ異なるプロットを使うのか？

(3) 独立な残差

データ収集順序に対する残差の散布図：

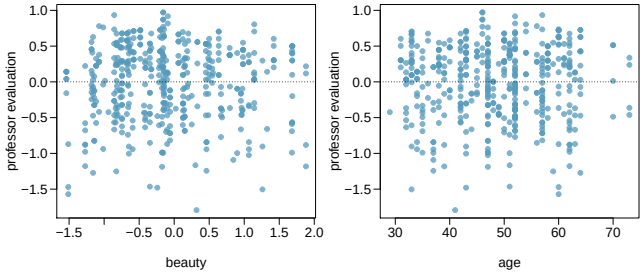
この条件は満たされているように見えるか？

独立な残差の条件についてさらに

- 残差が独立であることを確認することで、間接的に観測値の独立性を確認できる。
- 観測値と残差が独立であれば、データ収集順序に対する残差の散布図に増加または減少トレンドは見られないはずである。
- この条件は時系列データでしばしば満たされない。そのようなデータには適切な分析のために、より高度な時系列回帰技法が必要になる。

(4) 線形関係

各（数値）説明変数に対する残差の散布図：



この条件は満たされているように見えるか？

注：y 軸に予測変数の代わりに残差を使うのは、予測変数間の多重共線性などの他の違反を気にせずに線形性を確認できるからである。

モデルを改善するためのいくつかの方法

- 変数の変換
- モデルのギャップを埋める追加変数を探す
- 分散の不一致や予測変数と応答変数の非線形関係などの課題に対処できる、より高度な手法を使用する

変換

モデルの問題が説明変数と応答変数の間の非線形関係である場合、応答変数を変換することが有効なことがある。

- 対数変換（$\log y$）
- 平方根変換（$\sqrt{y}$）
- 逆数変換（$1/y$）
- 切り捨て（最大値を制限する）

説明変数に変換を適用することも可能だが、そのような変換はモデル係数の解釈をさらに困難にする傾向がある。

モデルは間違っていても役に立つ

すべてのモデルは間違っているが、いくつかは役に立つ。
— *George Box*

- 完璧なモデルは存在しないが、不完全なモデルでもモデルの欠点を明確に報告する限り役に立つ可能性がある。
- 条件が大きく違反している場合はモデルの結果を報告すべきではなく、より多くの統計的手法を学ぶか、専門家の助けを借りてでも新しいモデルを検討すべきである。

これまでの回帰 ...

ここまでに扱った内容：

- 単純線形回帰
  - 数値応答変数と数値またはカテゴリ予測変数の関係
- 重回帰
  - 数値応答変数と複数の数値およびカテゴリ予測変数の関係

まだ扱っていないのは、予測変数が変わった形（非線形、複雑な依存構造など）であったり、応答変数が変わった形（カテゴリ変数、カウントデータなど）であったりする場合の対処法である。

オッズ

オッズは事象の確率を表す別の方法であり、ギャンブル（およびロジスティック回帰）でよく使われる。

オッズ

ある事象 E に対して、

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

同様に、E のオッズが x 対 y であると言われたなら

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

これは次のことを意味する

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

例 — ドナー隊

1846 年、ドナー家とリード家はイリノイ州スプリングフィールドをカバーワゴンでカリフォルニアに向けて出発した。7 月、後に「ドナー隊」として知られることになるこの一行はワイオミング州フォートブリジャーに到達した。そこでリーダーたちはサクラメントバレーへの新しい未開拓ルートを試みることを決定した。87 人と 20 台のワゴンという最大規模に達した一行は、ワサッチ山脈の困難な横断と、グレートソルト湖西側の砂漠横断で再び遅延した。10 月下旬に激しい降雪がこの地域を襲い、グループはシエラネバダ山脈東部に立ち往生した。最後の生存者が 1847 年 4 月 21 日に救助されるまでに、87 人のうち 40 人が飢餓と極寒により死亡した。

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

例 — ドナー隊 — データ

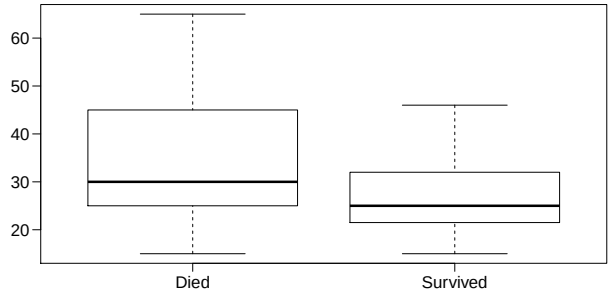
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

例 — ドナー隊 — 探索的データ分析

生存状況 vs. 性別：

	Male	Female
Died	20	5
Survived	10	10

生存状況 vs. 年齢：



例 — ドナー隊

年齢と性別の両方が生存に影響を与えることは明らかだが、この関係を探るためのモデルをどのように作ればよいか？

死亡を 0、生存を 1 に設定しても、変換で解決できる問題ではない — より高度な手法が必要である。

問題の捉え方の 1 つとして — 生存と死亡を、確率が線形モデルの変換で与えられる二項分布から生じる成功と失敗として扱うことができる。

一般化線形モデル

これは回帰におけるこの種の問題に対処する非常に一般的な方法であり、結果として得られるモデルを一般化線形モデル（GLM）と呼ぶ。ロジスティック回帰はこのタイプのモデルの一例に過ぎない。

すべての一般化線形モデルは以下の 3 つの特性を持つ：

1. 応答変数を記述する確率分布
2. 線形モデル
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$
3. 線形モデルと応答分布のパラメータを結びつけるリンク関数
   - $g(p) = \eta$ または $p = g^{-1}(\eta)$

ロジスティック回帰

ロジスティック回帰は、数値変数とカテゴリ変数の予測変数を用いて 2 値のカテゴリ応答変数をモデル化するための GLM である。

応答変数が二項分布から生じると仮定し、所与の予測変数セットに対する成功確率 p をモデル化したい。

ロジスティックモデルの定式化を完成させるために、η と p を結びつける適切なリンク関数を設定する必要がある。様々な選択肢があるが、最もよく使われるのはロジット関数である。

ロジット関数

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ ただし } 0 \leq p \leq 1$$

ロジットの性質

ロジット関数は 0 から 1 の間の値を取り、$-\infty$ から ∞ の値にマッピングする。

逆ロジット（ロジスティック）関数

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$

逆ロジット関数は $-\infty$ から ∞ の間の値を取り、0 から 1 の値にマッピングする。

この定式化はモデルの解釈にも有用で、ロジットは成功の対数オッズとして解釈できる（詳細は後述）。

ロジスティック回帰モデル

GLM の 3 つの基準から得られる：

$$y_i \sim \mathsf{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\mathsf{logit}(p) = \eta$$

ここから次の式が導かれる：

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1+\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

例 — ドナー隊 — モデル

R では、`lm` の代わりに `glm` を使い、`family` 引数で適合する GLM の種類を指定することで、線形モデルと同様の方法で GLM を適合できる。

```
summary(glm(Status ~ Age, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.81852    0.99937   1.820   0.0688 .
## Age         -0.06647    0.03222  -2.063   0.0391 *
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 56.291  on 43  degrees of freedom
## AIC: 60.291
##
## Number of Fisher Scoring iterations: 4
```


例 — ドナー隊 — 予測

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

モデル：

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \mathsf{Age}$$

新生児（Age=0）の生存オッズ/確率：

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$
$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$
$$p = 6.16/7.16 = 0.86$$

例 — ドナー隊 — 予測（続き）

モデル：

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

25 歳の生存オッズ/確率：

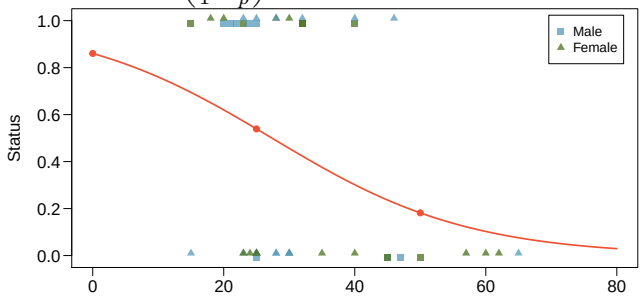
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$
$$\frac{p}{1-p} = \exp(0.156) = 1.17$$
$$p = 1.17/2.17 = 0.539$$

50 歳の生存オッズ/確率：

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50$$
$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$
$$p = 0.222/1.222 = 0.181$$

例 — ドナー隊 — 予測（続き）

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



例 — ドナー隊 — 解釈

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

単純な解釈は切片と傾きの項について、対数オッズと対数オッズ比の観点でのみ可能である。

切片：年齢 0 のメンバーの生存の対数オッズ。ここからオッズや確率を計算できるが、追加計算が必要である。

傾き：年齢が 1 単位増加する（1 歳年上になる）と対数オッズ比がどれだけ変化するか。直感的には分かりにくい。多くの場合、符号と相対的な大きさのみを気にする。

例 — ドナー隊 — 解釈 — 傾き

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.8185 - 0.0665(x+1)$$
$$= 1.8185 - 0.0665x - 0.0665$$
$$\log\left(\frac{p_2}{1-p_2}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = -0.0665$$
$$\log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}\right) = -0.0665$$
$$\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2} = \exp(-0.0665) = 0.94$$

例 — ドナー隊 — 年齢と性別

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

性別の傾き：他の予測変数を一定に保つとき、これは与えられた水準（Female）と参照水準（Male）の対数オッズ比である。

例 — ドナー隊 — 性別別モデル

重回帰と同様に、性別を代入することで男女それぞれの生存状況 vs. 年齢のモデルを得ることができる。

一般モデル：

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

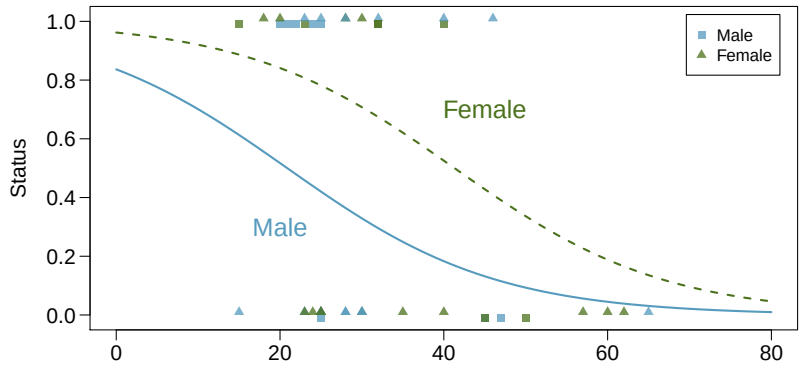
男性モデル：

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 0 \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

女性モデル：

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 1 \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

例 — ドナー隊 — 性別別モデル（続き）



モデル全体の仮説検定

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

注：モデル出力には F 統計量が含まれていない。一般的なルールとして、*GLM* モデルにはモデル全体の仮説検定が存在しない。

係数の仮説検定

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

ただし、個々の係数についての推測は依然として可能であり、基本的な設定はこれまでに見てきたものとまったく同じだが、Z 検定を使用する。

注：唯一の難点は、このコースの範囲をはるかに超えるが、標準誤差の計算方法である。

年齢の傾き係数の検定

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	*-0.0782*	*0.0373*	*-2.10*	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$
$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta_{age}} - \beta_{age}}{SE_{age}} = \frac{\textit{-0.0782} - 0}{\textit{0.0373}} = \textit{-2.10}$$

$$\text{p-value} = P(|Z| > \textit{2.10}) = P(Z > \textit{2.10}) + P(Z < \textit{-2.10})$$
$$= 2 \times 0.0178 = 0.0359$$

年齢の傾き係数の信頼区間

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

傾きの解釈は、予測変数の 1 単位の変化に対する対数オッズ比の変化であることを覚えておく。

対数オッズ比：

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

オッズ比：

$$\exp(CI) = (\exp -0.1513, \exp -0.0051) = (0.8596,\ 0.9949)$$

例 — 野鳥飼育と肺がん

1972〜1981 年にオランダのハーグで実施された健康調査により、ペットの鳥の飼育と肺がんリスクの増加との関連が発見された。鳥の飼育をリスク因子として調査するために、研究者らは 1985 年にハーグ（人口 45 万人）の 4 つの病院で症例対照研究を実施した。彼らは一般開業医に登録されており、65 歳以下で、1965 年以降ハーグに居住していた患者の中から 49 例の肺がん症例を特定した。また、同じ年齢構成を持つ住民から 98 人の対照を選んだ。

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

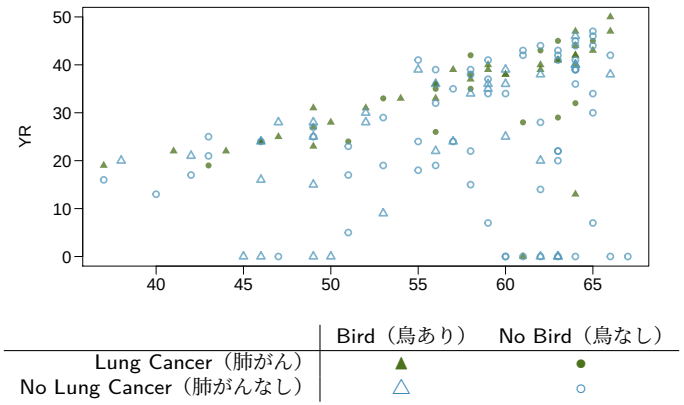
例 — 野鳥飼育と肺がん — データ

	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37.00	19.00	12.00
2	LungCancer	Male	Low	Bird	41.00	22.00	15.00
3	LungCancer	Male	High	NoBird	43.00	19.00	15.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
147	NoCancer	Female	Low	NoBird	65.00	7.00	2.00

LC 肺がんの有無
FM 対象者の性別
SS 社会経済的状況
BK 鳥の飼育の有無
AG 対象者の年齢（歳）
YR 診断または検査前の喫煙年数
CD 平均喫煙率（1 日当たりの本数）

注：*NoCancer* が参照応答（*0* または失敗）、*LungCancer* が非参照応答（*1* または成功）— これは解釈に重要である。

例 — 野鳥飼育と肺がん — 探索的データ分析



例 — 野鳥飼育と肺がん — モデル

```
summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))

## Call:
## glm(formula = LC ~ FM + SS + BK + AG + YR + CD, family = binomial,
##     data = bird)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736    1.80425  -1.074 0.282924
## FMFemale     0.56127    0.53116   1.057 0.290653
## SSHigh       0.10545    0.46885   0.225 0.822050
## BKBird       1.36259    0.41128   3.313 0.000923 ***
## AG          -0.03976    0.03548  -1.120 0.262503
## YR           0.07287    0.02649   2.751 0.005940 **
## CD           0.02602    0.02552   1.019 0.308055
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
##
## Number of Fisher Scoring iterations: 5
```


例 — 野鳥飼育と肺がん — 解釈

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

他のすべての予測変数を一定に保つと、

- 鳥を飼育している人と飼育していない人の肺がんに関するオッズ比は $\exp(1.3626) = 3.91$ である。
- 喫煙年数が 1 年増加することによる肺がんのオッズ比は $\exp(0.0729) = 1.08$ である。

数値が意味しないこと ...

ロジスティック回帰の解釈でよくある間違いは、オッズ比を確率の比として扱うことである。

鳥を飼育している人は、飼育していない人より肺がんを発症する可能性が 4 倍高いわけではない。

これはリスク比とオッズ比の違いである。

$$RR = \frac{P(\text{疾患}|\text{曝露あり})}{P(\text{疾患}|\text{曝露なし})}$$

$$OR = \frac{P(\text{疾患}|\text{曝露あり})/[1 - P(\text{疾患}|\text{曝露あり})]}{P(\text{疾患}|\text{曝露なし})/[1 - P(\text{疾患}|\text{曝露なし})]}$$

鳥の話に戻る

$P(\text{肺がん}|\text{鳥なし}) = 0.05$ とわかっている場合、鳥を飼育している人の肺がん確率は？

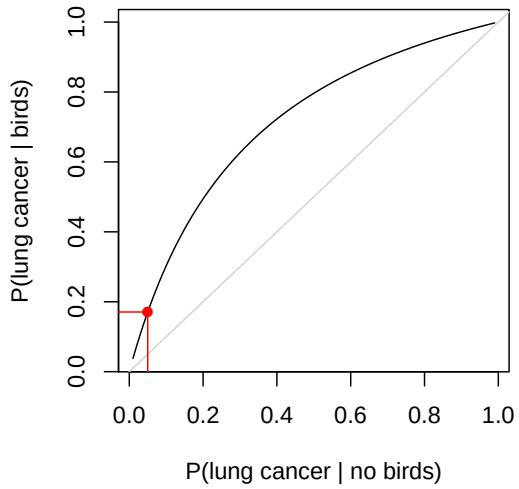
$$OR = \frac{P(\text{肺がん}|\text{鳥あり})/[1 - P(\text{肺がん}|\text{鳥あり})]}{P(\text{肺がん}|\text{鳥なし})/[1 - P(\text{肺がん}|\text{鳥なし})]}$$

$$= \frac{P(\text{肺がん}|\text{鳥あり})/[1 - P(\text{肺がん}|\text{鳥あり})]}{0.05/[1 - 0.05]} = 3.91$$

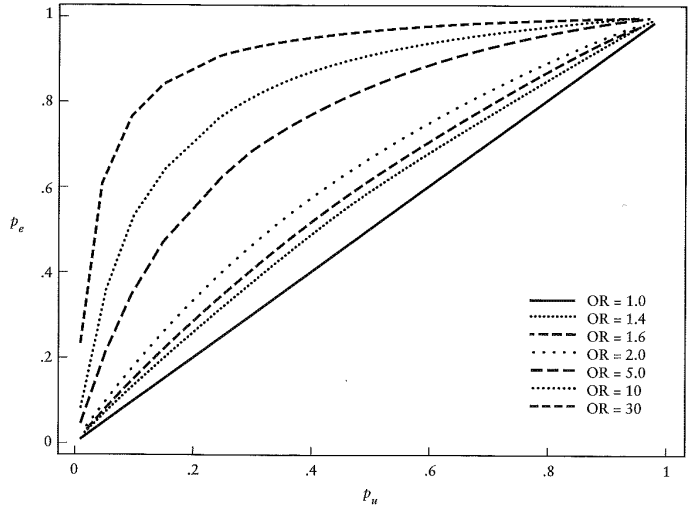
$$P(\text{肺がん}|\text{鳥あり}) = \frac{3.91 \times \frac{0.05}{0.95}}{1 + 3.91 \times \frac{0.05}{0.95}} = 0.171$$

$$RR = P(\text{肺がん}|\text{鳥あり})/P(\text{肺がん}|\text{鳥なし}) = 0.171/0.05 = 3.41$$

鳥の OR 曲線



OR 曲線



（昔の）例 — House

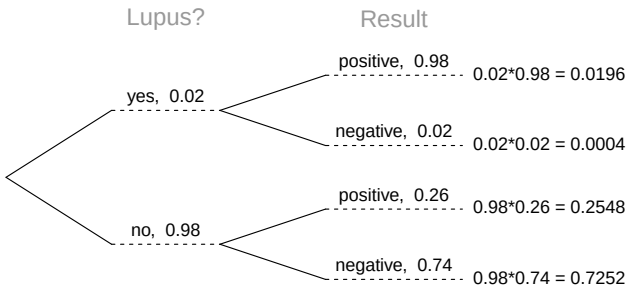
Fox のテレビ番組 House を見たことがあれば、House 医師が「ループスであることはない」とよく言うのを知っているだろう。

ループスは、外来細胞を攻撃して感染を防ぐはずの抗体が代わりに血漿タンパク質を外来物として認識し、血栓のリスクが高くなる医学的現象である。人口の 2% がこの疾患に罹患していると考えられている。

ループスの検査は、実際にループスがある場合は非常に正確だが、ない場合は非常に不正確である。より具体的には、実際に疾患がある人に対して 98% の精度を持つ。疾患がない人に対しては 74% の精度を持つ。

ループスの検査が陽性であっても House 医師は正しいか？

（昔の）例 — House



$$P(\text{ループス}|+) = \frac{P(+, \text{ループス})}{P(+, \text{ループス}) + P(+, \text{ループスなし})}$$
$$= \frac{0.0196}{0.0196 + 0.2548} = 0.0714$$

ループスの検査

ループスの検査は実際には非常に複雑で、診断は通常、複数の検査結果に基づく。検査には全血球計算、赤血球沈降速度、腎臓・肝臓評価、尿検査、抗核抗体（ANA）検査などが含まれることが多い。

これらの各検査に何が関わるか（例えば全血球計算が高いか低いかの判断）と、個々の検査および関連する判断が患者へのループス診断の全体的な決定においてどのような役割を果たすかを考えることが重要である。

ループスの検査

ある意味で、診断は様々な説明変数の複雑な統合を伴う二値決定（ループスかループスでないか）と見ることができる。

この例は診断がどのように行われるかについての情報を与えてくれないが、与えてくれる同様に重要なことがある — それが検査の感度と特異度である。これらの値は陽性または陰性の検査結果が実際に何を意味するかを理解するために重要である。

感度と特異度

感度（*Sensitivity*）— 検査が陽性結果を正しく識別する能力を測定する。

$$P(\text{検査}+ \mid \text{疾患}+) = P(+|\text{ループス}) = 0.98$$

特異度（*Specificity*）— 検査が陰性結果を正しく識別する能力を測定する。

$$P(\text{検査}- \mid \text{疾患}-) = P(-|\text{ループスなし}) = 0.74$$

極端なケースについて考えると理解が深まる — 常に陽性を返す検査の感度と特異度は？ 常に陰性を返す検査は？

感度と特異度（続き）

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\text{感度} = P(\text{検査}+ \mid \text{疾患}+) = TP/(TP+FN)$$
$$\text{特異度} = P(\text{検査}- \mid \text{疾患}-) = TN/(FP+TN)$$
$$\text{偽陰性率 } (\beta) = P(\text{検査}- \mid \text{疾患}+) = FN/(TP+FN)$$
$$\text{偽陽性率 } (\alpha) = P(\text{検査}+ \mid \text{疾患}-) = FP/(FP+TN)$$

$$\text{感度} = 1 - \text{偽陰性率} = \text{検出力}$$
$$\text{特異度} = 1 - \text{偽陽性率}$$

それで？

検査の感度と特異度（および/または偽陽性率と偽陰性率）を知ることが重要であることは明らかである。疾患の有病率（例：$P(\text{ループス})$）とともに、これらの値は $P(\text{ループス}|+)$ のような重要な量を計算するために必要である。

また、第 1 回中間試験前の検出力分析への簡単な取り組みも、偽陽性率と偽陰性率を最小化することに伴うトレードオフ（検出力を高めるには α または n を増やす必要がある）について示唆を与えているはずである。

意思決定を行おうとするとき、この情報をどのように使うべきか？

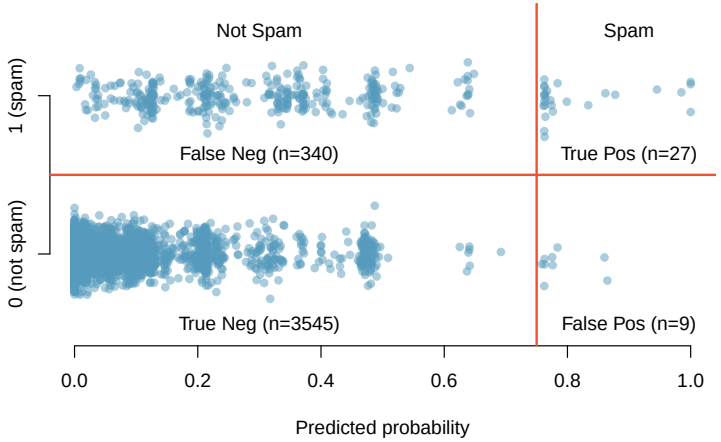
スパムに戻る

今週の実験では、スパムメッセージを識別することに関心があるメールのデータセットを調べた。異なる予測変数がメッセージがスパムである確率にどう影響するかを評価するために、異なるロジスティック回帰モデルを検討した。

これらのモデルは受信メッセージに確率を割り当てるためにも使える（これは SLR / MLR の場合の予測に相当する）。しかし、スパムフィルターを設計するとしたら、これは半分の戦いに過ぎない。これらの確率を使って、どのメールをスパムとしてフラグを立てるかについての決定も行う必要がある。

唯一の解決策ではないが、閾値確率を選択し、その確率を超えたメールはスパムとしてフラグを立てるという単純なアプローチを考える。

閾値の選択



閾値を *0.75* に設定した場合を見てみよう。

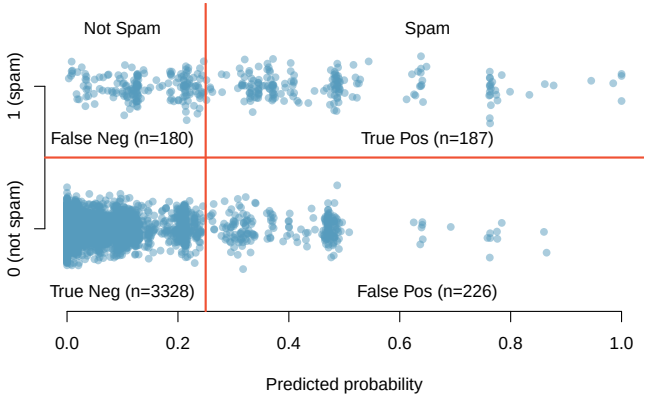
閾値選択の結果

メールデータセットで閾値 0.75 を選ぶと以下の結果が得られる：

$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

この意思決定ルールの感度と特異度は？

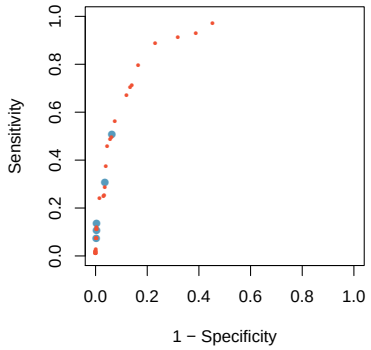
他の閾値を試す



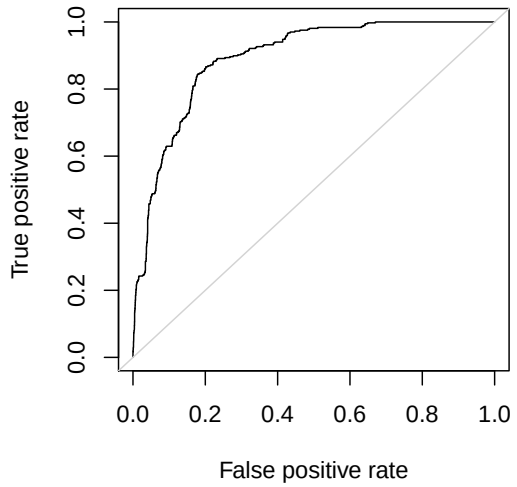
閾値	0.75	0.625	0.5	0.375	0.25
感度	0.074	0.106	0.136	0.305	0.510
特異度	0.997	0.995	0.995	0.963	0.936

感度と特異度の関係

閾値	0.75	0.625	0.5	0.375	0.25
感度	0.074	0.106	0.136	0.305	0.510
特異度	0.997	0.995	0.995	0.963	0.936



受信者操作特性（ROC）曲線



受信者操作特性（ROC）曲線（続き）

なぜ ROC 曲線を使うのか？

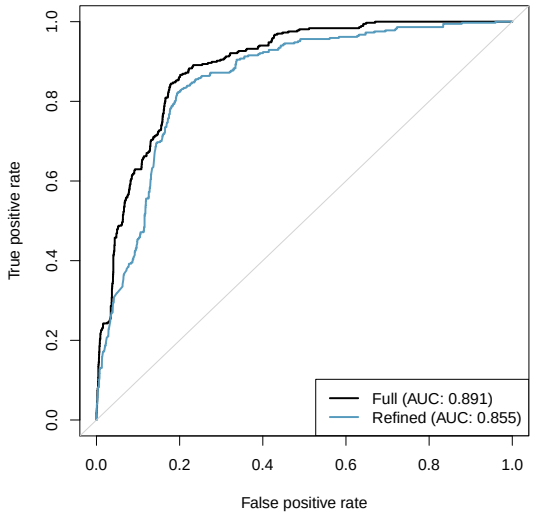
- すべての可能な閾値における感度と特異度のトレードオフを示す。
- 偶然による性能との比較が直感的にわかる。
- 曲線下面積（AUC）をモデルの予測能力の評価として使用できる。

スパムモデルの精緻化

```
g_refined = glm(spam ~ to_multiple+cc+image+attach+winner
                      +password+line_breaks+format+re_subj
                      +urgent_subj+exclaim_mess,
                data=email, family=binomial)
summary(g_refined)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7594	0.1177	-14.94	0.0000
to_multipleyes	-2.7368	0.3156	-8.67	0.0000
ccyes	-0.5358	0.3143	-1.71	0.0882
imageyes	-1.8585	0.7701	-2.41	0.0158
attachyes	1.2002	0.2391	5.02	0.0000
winneryes	2.0433	0.3528	5.79	0.0000
passwordyes	-1.5618	0.5354	-2.92	0.0035
line_breaks	-0.0031	0.0005	-6.33	0.0000
formatPlain	1.0130	0.1380	7.34	0.0000
re_subjyes	-2.9935	0.3778	-7.92	0.0000
urgent_subjyes	3.8830	1.0054	3.86	0.0001
exclaim_mess	0.0093	0.0016	5.71	0.0000

モデルの比較



効用関数

「最良の」閾値を決定するために使える合理的な定量的アプローチは他にも多くある。

経済学のコースを受けたことがあれば、効用関数のアイデアを聞いたことがあるだろう。各結果に対してコストと利益を割り当て、各状況の効用を計算するためにそれらを使うことができる。

スパムフィルターの効用関数

スパムフィルターの効用関数を記述するには、各結果のコスト/利益を考慮する必要がある。

結果	効用
真陽性（TP）	1
真陰性（TN）	1
偽陽性（FP）	-50
偽陰性（FN）	-5

$$U(p) = TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p)$$

閾値 0.75 の効用

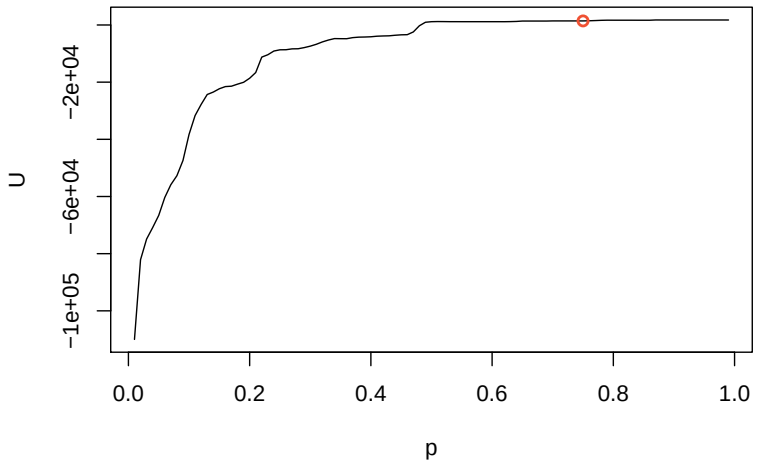
メールデータセットで閾値 0.75 を選ぶと以下の結果が得られる：

$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

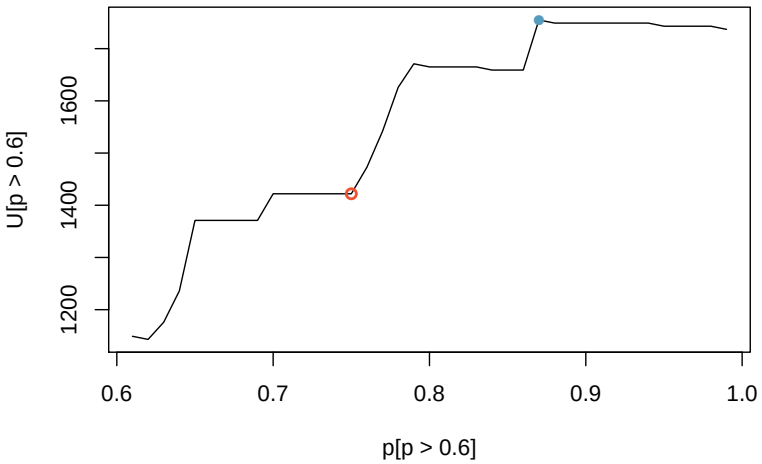
$$\begin{aligned} U(p) &= TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p) \\ &= 27 + 3545 - 50 \times 9 - 5 \times 340 = 1422 \end{aligned}$$

それ自体では有用ではないが、他の閾値と比較することができる。

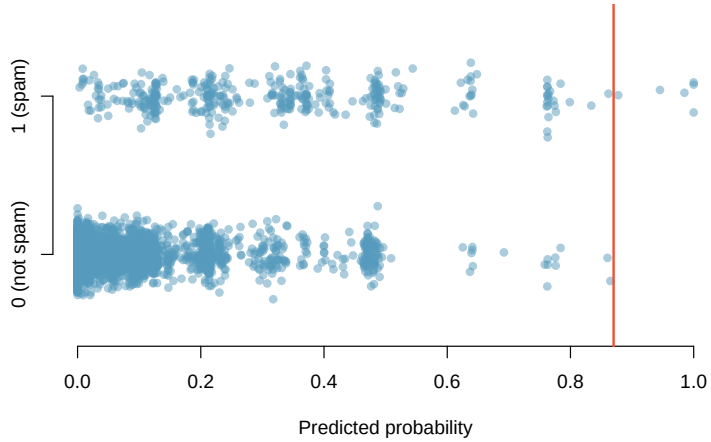
効用曲線



効用曲線（拡大）


U[p > 0.6]
1200
1400
1600
0.6
0.7
0.8
0.9
1.0
p[p > 0.6]

最大効用


1 (spam)
0 (not spam)
0.0
0.2
0.4
0.6
0.8
1.0
Predicted probability