政治学方法論I課題11

提出者：宋財泫(ソンジェヒョン)(123J009J)　提出日：2014-12-19(金)

1　問題 1

確率変数 Y_i が互いに独立で、

$$Y_i \sim \text{Bin}(n, \pi_i)$$
$$\pi_i = \text{logit}^{-1}(\beta_1 + \beta_2 x_i)$$

というモデルを想定し、表 1 のデータを使ってロジスティック回帰を実行しなさい。ただし、4 以外は glm() 関数は使わずに、自分で定義した対数尤度関数を使いなさい（最尤推定値の探索には、maxLik() を使ってかまわない）。

表 1

x_i	n_i	y_i
5	8	0
10	5	0
15	8	4
20	10	2
25	10	7
30	6	3
35	3	1
40	5	5
45	4	4
50	4	4

問 1-1　このデータに対する θ の対数尤度関数を図示しなさい。ただし、β_1 を 1 つの値に固定し、横軸に β_2 をとること。

線形予測子の切片 (β_1) を-1 に固定し、パラメータ β_2 を-0.5 から 0.5 まで 0.0001 単位で変化させながら対数尤度の変化を見ると $\beta_2 = 0.0523$ の地点で対数尤度は最大となり、最大対数尤度は-34.87273 である。これらの結果は図 1 で確認できる。

(図 1、 2 この辺り)

また、後述する問題 1-2 から得られた $\hat{\beta}_1$、-3.361014 で β_1 を固定し、β_2 を 0 から 0.5 まで 0.0001 単位で変化させた場合の対数尤度関数は $\hat{\beta}_2 = 0.136604$ で最大値-29.66783 が得られることが図 2 から確認出来る。

問 1-2　最尤推定値 (MLE) を求めなさい。

最尤推定値を求めるための関数を作成し、maxLik() 関数を用いて推定を行う。推定方法としてはニュートン–ラフソン法、初期値は $\beta_1 = -3, \beta_2 = 0.15$ である。これらの過程より得られた推定値は以下のようである。

```
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 5 iterations
Return code 1: gradient close to zero
Log-Likelihood: -29.66783
2  free parameters
Estimates:
      Estimate Std. error t value   Pr(> t)
[1,] -3.361014   0.874330 -3.8441  0.000121
[2,]  0.136604   0.034878  3.9166 8.979e-05
--------------------------------------------
```

以上の結果から β_1 は-3.361014、β_2 は 0.136604 という推定値が得られた。

問 1-3　AIC を求めなさい。

赤池情報量基準 (Akaike's Information Criterion; AIC) は $-2\times 2\text{LogLikelihood}\times 2*k$ という式で求めることができる。k は問 1-2 の推定結果における free parameters を意味し、したがって AIC は $-2 \times 2 \times -29.66783 + 2 \times 2 = 63.33566$ である。

問 1-4　glm() を使い、MLE と AIC を求め、上と同じ結果が得られるか確認しなさい。

最初に表 1 のデータの二項分布のデータで変形し、つづいて R の glm 関数を用いてロジスティクス回帰分析を行った。推定の結果は以下のようである。

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.36101    0.87343  -3.848 0.000119
new.x        0.13660    0.03485   3.920 8.85e-05

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 87.194  on 62  degrees of freedom
Residual deviance: 59.336  on 61  degrees of freedom
AIC: 63.336

Number of Fisher Scoring iterations: 5
```

glm 関数による結果と問題 1-3 の推定値と比較すると

(表 2 この辺り)

であり、推定値はほぼ一致するものの、標準誤差の場合、小数点三桁目からややズレが確認できる。

2　問題 2

授業の web 資料「ロジスティック回帰 (2)」で説明されている架空の選挙データを使い、過去の当選回数と選挙費用という 2 つの説明変数で選挙の当落を説明するロジスティック回帰分析を実行しなさい。

問 2-1　自分で対数尤度関数を定義し、glm() を使わずに、最尤推定値と AIC を求めなさい。

問題 1 とは異なり、この場合は二項分布ではなくベルヌーイ分布であるため、対数尤度関数は

$$\pi_i = \frac{1}{1-\exp(-(\beta_{1,i}+\beta_2 Previous_i+\beta_3 Expm_i))}$$
$$\text{Likelihood} = \prod_{i=1}^{N} \pi_i^{\text{wlsmd}_i}(1-\pi_i)^{1-\text{wlsmd}_i}$$
$$\text{LogLikelihood} = \sum_{i=1}^{N} \log\pi_i^{\text{wlsmd}_i}(1-\pi_i)^{1-\text{wlsmd}_i} \tag{1}$$

であり、この関数を R の maxLik 関数を用いて β_1、β_2、β_3 を推定する。推定結果は

```
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 7 iterations
Return code 1: gradient close to zero
Log-Likelihood: -5.192101
3  free parameters
Estimates:
     Estimate Std. error t value Pr(> t)
[1,] -6.38111    3.49402 -1.8263 0.06781
[2,]  0.80853    0.58263  1.3877 0.16522
[3,]  0.80882    0.39815  2.0315 0.04221
--------------------------------------------
```

である。具体的には $\beta_1 = -6.38111$、$\beta_2 = 0.80853$、$\beta_3 = 0.80882$ であり、AIC は $-2 \times -5.192101 + 2 \times 3 = 16.3842$ である。

問 2-2　glm() を使って最尤推定値と AIC を求め、1 と同様の結果が得られることを確認しなさい。

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.3811     3.5147  -1.816   0.0694
previous      0.8085     0.5851   1.382   0.1670
expm          0.8088     0.4000   2.022   0.0431

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20.728  on 14  degrees of freedom
Residual deviance: 10.384  on 12  degrees of freedom
AIC: 16.384

Number of Fisher Scoring iterations: 6
```

問 2-1 の推定結果との比較は以下の表 3 のようである。

(表 3 この辺り)

これらの結果は問 1-4 と同様、推定値と AIC は一致するが標準誤差がやや異なることが確認できる。

問 2-3　(1) 当選確率の予測値が 0.4 以上の候補者を当選と考える場合と、(2) 当選確率の予測値が 0.6 以上の候補者を当選と考える場合について、それぞれ予測の的中率を求めなさい。

```
> Prob2.3.pred <- predict(Prob2.2.glm, type="response")
> Cor.rate.1 <- mean((Prob2.3.pred >= 0.4 &
+                     Prob2.2.df$wlsmd == 1) |
+                    (Prob2.3.pred <0.4 &
+                     Prob2.2.df$wlsmd == 0))
> print(Cor.rate)
[1] 0.7333333
>
> Cor.rate.2 <- mean((Prob2.3.pred >= 0.6 &
+                     Prob2.2.df$wlsmd == 1) |
+                    (Prob2.3.pred <0.6 &
+                     Prob2.2.df$wlsmd == 0))
> print(Cor.rate.2)
[1] 0.8
```

当落の予測確率の基準が 0.4 の場合、的中率は 0.7333 であり、0.6 の場合は 0.8 である。これらの的中率は R を用いて簡単に計算する事ができ、表 4 のような直観的に理解することも出来る。

(表 4 この辺り)

問 2-4　ROC 曲線を描き、当てはまり具合を評価しなさい。

```
Call:
roc.formula(formula = Prob2.2.df$wlsmd ~ Prob2.3.pred,
            plot = TRUE)

Data: Prob2.3.pred in 7 controls (Prob2.2.df$wlsmd 0) <
      8 cases (Prob2.2.df$wlsmd 1).
Area under the curve: 0.8929
```

(図 3 この辺り)

総合的な指標としての AUC は 0.8929 であり、全体的には良いモデルだと考えられる。ROC 曲線、図 3 を見ると横軸 (Specificity) が 1.0 の時、つまりこのモデルにおいて応答変数が 0 のケースを 1 と予測した比率が 100% の場合においても的中率 (Sensitivity) はおよそ 70% であり、横軸がおよそ 0.5 の時には的中率が 100% まで増加することが確認できる。複数のモデルを比較しないかぎり、このモデルの相対的な良さは分からないが、ある程度、当てはまりの良いモデルであるとは評価できよう。

3　問題 3

課題 9 で集めたデータを使い、ロジスティック回帰分析を行いなさい。必要があれば、複数のモデルで分析すること。

課題 9 で提出した CSES Module 3 のデータ (日韓に限定) を用いて投票参加の仮説を検証する。

H1　年齢が高いほど投票する傾向がある。
H2　教育水準が高いほど投票する傾向がある。
H3　収入が高いほど投票する傾向がある。

モデルは 4 つのモデルを検証するが、モデル 1 は仮説 1 のみを、モデル 2 は仮説 1 と 2 を、モデル 3 は仮説 1、2、3 を、モデル 4 はモデル 3 に国ダミーを投入する。また、年齢は二次曲線的な影響力を持つと考えられるが、本課題では割愛し、また国ダミーと他の変数間の交互作用も検証しない。

分析に用いる変数は 5 つ (1 つの応答変数と 4 つの説明変数) であり、以下のように操作化を行った。元変数の情報は CSES のホームページのコードブックを参照すること。また、分析の前段階に欠損値のデータは listwise で全て除去し、最終的に残ったサンプルサイズは $N = 1,760$ である。

投票参加　元変数 `C3021_1` をそのまま利用。`DK/NA` は欠損値扱い
年齢　元変数 `C2001` をそのまま利用
教育　元変数 `C2003` をそのまま。`DK/NA` は欠損値扱い
所得　元変数 `C2020` をそのまま。`DK/NA` は欠損値扱い
国　日本 = 0; 韓国 = 1

上記のように操作化された変数の記述等計は以下の表 5 である。

(表 5 この辺り)

問 3-1　自分で対数尤度関数を定義し、glm() を使わずに、最尤推定値と AIC を求めなさい。

(推定過程は省略)

(推定結果の表 6 この辺り)

問 3-2　glm() を使って最尤推定値と AIC を求め、1 と同様の結果が得られることを確認しなさい。

(推定過程は省略)

(推定結果の表 7 この辺り)

glm 関数を用いた検証結果、表 7 で確認できるように係数の推定値は一致するが、標準誤差の推定値は小数点三桁目でズレがある変数がいくつかある。

問 3-3　予測確率が 0.5 以上の場合に応答変数が 1 をとると想定し、予測の的中率を求めなさい。

各モデルの的中率は

モデル 1　0.7670455 (76.70%)
モデル 2　0.7659091 (76.59%)
モデル 3　0.7698864 (76.98%)
モデル 4　0.7710227 (77.10%)

であり、モデル 2 では的中率の改善は確認できないが、モデル 3 と 4 においては若干の改善が確認できる。しかし、約 0.4 ポイントの改善であり、大きいとは言いがたい。これらの結果は概ね表 6 と表 7 の AIC の傾向と一致するがモデル 2 の場合、AIC から見ればモデルは改善されたものの、的中率を基準とすれば約 0.1 ポイント悪化した。むろん、両モデルにおける AIC の差は大きくないため納得できる傾向であろう。

問 3-4　ROC 曲線を描き、当てはまり具合を評価しなさい。

(図 4 から 7 この辺り)

各モデルの AUC は

モデル 1　0.6884
モデル 2　0.6904
モデル 3　0.7017
モデル 4　0.7309

であり、AIC と同様、モデルに変数が追加されるにつれてモデルがやや改善される傾向が確認できる。モデル 1 とモデル 4 の AUC の差は約 0.04 であり、これがどれだけモ

デルを改善したかを語ることは難しいが、改善の有無の面で言えば、より良いモデルとも言えよう。モデル 2 の場合は AIC と AUC から見れば良いモデルであろうが、的中率が若干下がったため積極的には主張できないが、少なくともモデル 3 とモデル 4 はモデル 1 と比べて大小に関係なく良いモデルだと言えよう。

4　図表

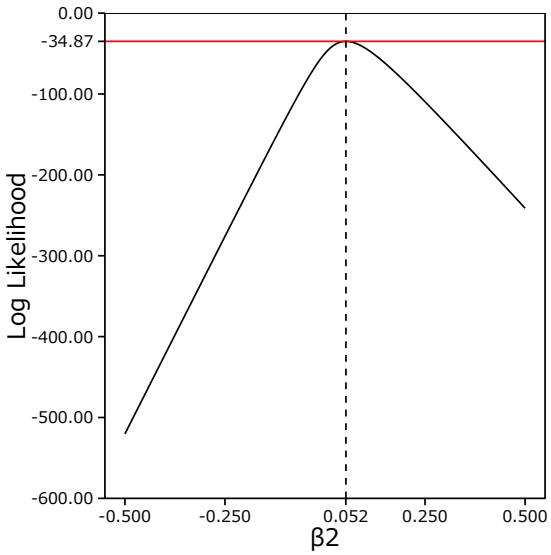

図 1: 問 1-1(1)

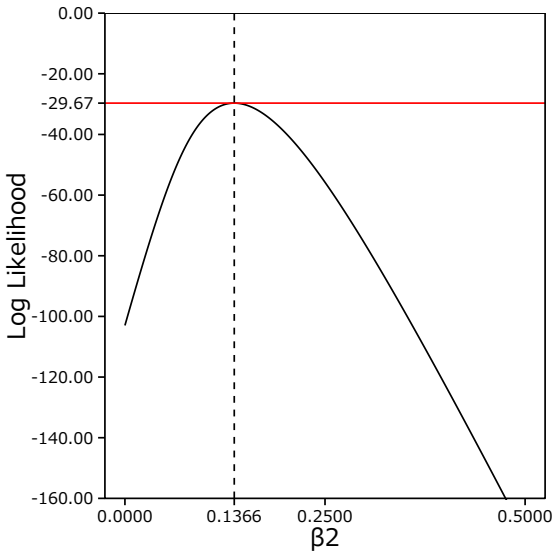

図 2: 問 1-1(2)

表 2: 自作関数 +maxLik の推定値と glm による推定値の比較

	問 1-3		問 1-4		Δ	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
x	0.13660	0.03479	0.13660	0.03485	0.00000	0.00006
Intercept	-3.36101	0.87172	-3.36101	0.87343	0.00000	0.00171
AIC	63.336		63.336		0.000	
注: Δ は問 1-3 と問 1-4 の推定値の差						

表 3: 自作関数 +maxLik の推定値と glm による推定値の比較 (2)

	問 2-1		問 2-2		Δ	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Previous	0.8085	0.5826	0.8085	0.5851	0.0000	0.0025
Expm	0.8088	0.3982	0.8088	0.4000	0.0000	0.0018
Intercept	-6.3811	3.4940	-6.3811	3.5147	0.0000	0.0207
AIC	16.384		16.384		0.000	
注: Δ は問 2-1 と問 2-2 の推定値の差						

表 4: 的中率の比較

ID	当落	予測確率	当落予測 (0.4)	当落予測 (0.6)	的中 (0.4)	的中 (0.6)
1	1	0.846	1	1	○	○
2	1	0.925	1	1	○	○
3	1	0.962	1	1	○	○
4	1	0.980	1	1	○	○
5	1	0.975	1	1	○	○
6	1	0.327	0	0	×	×
7	1	0.327	0	0	×	×
8	1	0.925	1	1	○	○
9	0	0.010	0	0	○	○
10	0	0.052	0	0	○	○
11	0	0.710	1	1	×	×
12	0	0.088	0	0	○	○
13	0	0.522	1	0	×	○
14	0	0.022	0	0	○	○
15	0	0.327	0	0	○	○

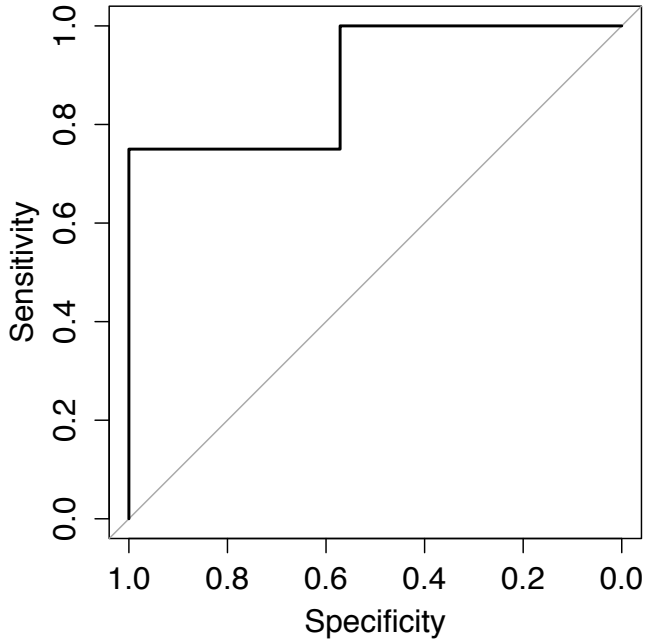

図 3: 問 2-4 の ROC プロット

表 5: 問題 3 の記述統計

	Mean	SD	Min	Max	
vote			0	1	棄権 = 405; 投票 = 1355
age	49.23	15.74	19	91	
educ			1	8	1 = 14; 2 = 19; 3 = 171; 4 = 97
					5 = 693; 6 = 173; 7 = 139; 8 = 454
income			1	5	1 = 282; 2 = 469; 3 = 590; 4 = 271
					5 = 148
korea			0	1	日本 (0) = 944; 韓国 (1) = 816

表 6: CSES Module 3 を用いた日韓の投票参加の仮説検証結果

	モデル 1	モデル 2	モデル 3	モデル 4
年齢	0.044***	0.048***	0.049***	0.041***
	(0.004)	(0.005)	(0.005)	(0.005)
教育		0.065	0.022	0.047
		(0.041)	(0.042)	(0.043)
所得			0.246***	0.205***
			(0.058)	(0.062)
国ダミー [1)]				-0.942***
				(0.130)
切片	-0.856***	-1.394***	-1.859***	-0.992***
	(0.187)	(0.383)	(0.401)	(0.434)
AIC	1766.803	1766.264	1749.741	1697.311

注: $^{\dagger} \leq 0.1$;* ≤ 0.05;** ≤ 0.01;*** ≤ 0.001

[1)] ベースグループは日本 (0)

表 7: CSES Module 3 を用いた日韓の投票参加の仮説検証結果 (glm)

	モデル 1	モデル 2	モデル 3	モデル 4
年齢	0.044***	0.048***	0.049***	0.041***
	(0.004)	(0.005)	(0.005)	(0.005)
Δ_1	0.000	0.000	0.000	0.000
Δ_2	0.000	0.000	0.000	0.000
教育		0.065	0.022	0.047
		(0.041)	(0.043)	(0.043)
Δ_1		0.000	0.000	0.000
Δ_2		0.000	0.001	0.000
所得			0.246***	0.205***
			(0.058)	(0.062)
Δ_1			0.000	0.000
Δ_2			0.000	0.000
国ダミー [1)]				-0.942***
				(0.130)
Δ_1				0.000
Δ_2				0.000
切片	-0.856***	-1.394***	-1.859***	-0.992***
	(0.186)	(0.387)	(0.406)	(0.430)
Δ_1	0.000	0.000	0.000	0.000
Δ_2	0.000	0.004	0.005	0.004
AIC	1766.803	1766.264	1749.741	1697.311
Δ_1	0.000	0.000	0.000	0.000

注 1: $^{\dagger} \leq 0.1$;* ≤ 0.05;** ≤ 0.01;*** ≤ 0.001

注 2: Δ_1 は表 6 の推定値との差 (Coef.)

注 3: Δ_2 は表 6 の推定値との差 (SE)

[1)] ベースグループは日本 (0)

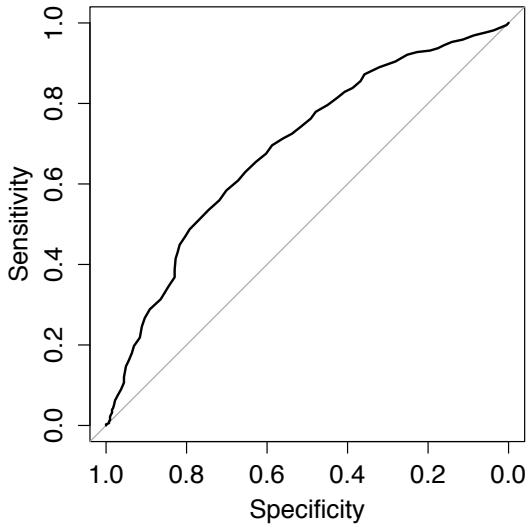

図 4: モデル 1 の ROC 曲線

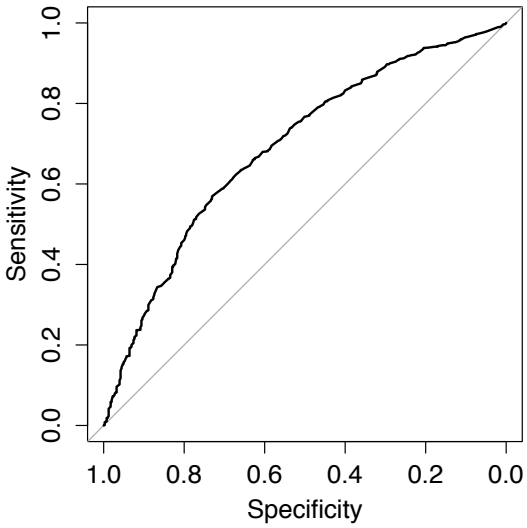

図 5: モデル 2 の ROC 曲線

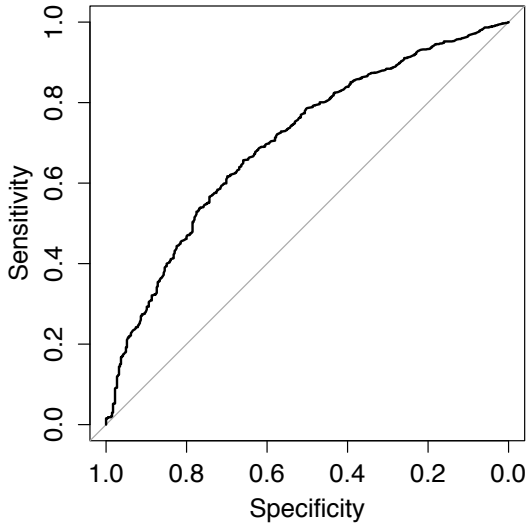

図 6: モデル 3 の ROC 曲線

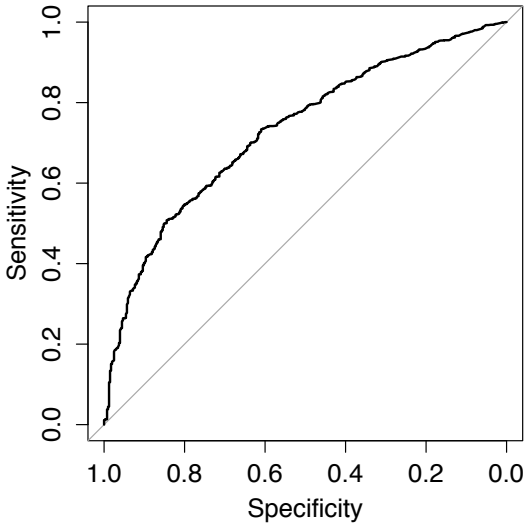

図 7: モデル 4 の ROC 曲線