政治学方法論Ⅰ課題9

提出者：宋財泫(ソンジェヒョン)(123J009J)　提出日：2014-12-06(土)

1　問題 1

以下の指示に従って、ロジスティック回帰分析のシミュレーションを行いなさい。

1. 自分で母数 (parameters) とデータ生成過程を設定し、データを生成しなさい。

```r
#Setting parameters
intercept <- 0.2 #Intercept
beta1 <- 0.14 #Coefficient of x
n <- 500 #Sample size

#Sampling n times from unifrom distribution which has a range from -10 to
    10
x <- runif(n, -10, 10)
#Calculating the probability with x and intercept
pr <- 1 / (1 + exp(-(intercept + beta1*x)))
#Sampling from binomial distribution with parameters n and pr
y <- rbinom(n = n, size = 1, prob = pr)
```

2. 生成したデータにロジスティック回帰を当てはめ、母数を推定しなさい。

(表 1)

3. 生成したデータに OLS を当てはめ、母数を推定しなさい。

(表 2)

4. 重回帰分析のシミュレーションのときと同様に、上の過程を繰り返し実行し、ロジスティック回帰の特徴を調べなさい。その際、様々な条件を（一つずつ）変え、どのような条件で推定がうまく行えるか（行えないか）を明らかにしなさい。また、応答変数が二値のデータに OLS を当てはめるとどのような問題が生じるか（あるいはそもそも問題があるのか）調べなさい。

1.1　シミュレーションの過程

シミュレーション関数のソースコードは付録を参照

1.2　サンプルサイズを増やす

本節ではサンプルサイズとロジスティック回帰分析の推定値の関係を調べるために、サンプルサイズ n を 10, 30, 50, 100, 500, 1000 に変えながらシミュレーションを行う。いずれの場合もサンプル数は 100 である。

図 1 と 2 はシミュレーションの結果である[1]。青あるいは赤い線は 95% 信頼区間をあらわすため、ほとんどの推定値の信頼区間にパラメータが含まれることが確認できる。しかし、信頼区間の広さはサンプルサイズによって大きく変わり、これは標準誤差がサンプルサイズの影響を受けるためである。むしろ注目すべきところは信頼区間よりも推定値そのものである。左上の $n = 10$ の場合、推定値が比較的に母数から遠くかけ離れているに対して、$n = 1,000$ の場合は母数の周辺に密集していることが確認できる。

(図 1、 2)

以上の結果をより詳細に調べるために、最後はサンプルサイズ n を 10 から 500 まで 1 ずつ増やしながら 10 回の試行による推定値の平均値のプロットを示す。赤い実線は母数であり、黒い実線は推定値の平均値、網掛けの部分は 10 回の試行の中で最小値と最大値の範囲を表す。n が小さい時に推定が不可能なときは全試行が成功するまで振り返した。以下の図 3 と 4 はこのシミュレーションの結果である。

(図 3、 4)

黒い実線が途切れている箇所や最大–最小値の範囲が示されていないところはグラフの y 軸の範囲を超えるケースである[2]。切片の推定値は大きな意味を持たず、n が増えるにつれて安定していくものの、変動の幅 (網掛けの部分) はほぼ変わらない。これに対して説明変数の係数 β は切片の推定値と同じく n の増加によって安定していく傾向と、その変動の幅 (網掛けの部分) も小さくなっていくことが確認できる。おおよそ n が 100 からは安定し、これは Long(1997) の主張とも整合的である。ただし、パラメータが多くなると安定した推定値を得るために必要なサンプルサイズは大きくなり[3]、本課題においてシミュレーションは割愛する。

---

[1] ggplot2 の場合、現在プロットを回転した後にプロットごとに異なるスケールを指定することができないため、やむを得ず元のグラフを載せた

[2] 推定値の平均だけでいうと切片の場合は-78.93、β の場合は 19.41 のケースがあった。

[3] 同じく MLE で推定を行っても順序ロジスティクス回帰分析の場合はより多くのサンプルサイズが求められる。

1.3　パラメータを変える 1–切片

パラメータを変えるということはデータ生成過程内のパラメータを変えることを意味する。ロジステック関数は $Pr(y=1|x)=\frac{1}{1+e^{-(\alpha+\beta x)}}$ であり、線形予測子の $\alpha+\beta x$ の値が高くなるとデータ生成過程から 1 が得られる確率は高くなり、結果的には応答変数の分布が変わる。応答変数の分布によって推定の結果を確認するために $n=100$、$\beta=0.14$、試行回数は 100 で固定した上で切片のみを 0、0.1、0.5、1、5、10 に変化させて推定結果を比較する。まずは切片の変化が応答変数の分布に与える影響を確認する。

(表 3)

表 3 を見ると切片が大きくなるとデータ生成過程で 1 が得られる確率が増加し、10 の場合は全ての応答変数の値が 1 になるという、つまり応答変数に分散がない状態に陥る。応答変数に分散がない場合、いかなる推定も不可能である。したがって、ある程度の分散が保証されている $\alpha=0,0.1,0.5,1$ のケースのみでシミュレーションを行う。

(図 5、 6)

図 5 と 6 から応答変数の分散がある程度ある場合における推定値が読み取れる。応答変数の分布が極端に偏らないかぎり、推定値のは安定する。しかし、分布が極端に偏る場合はどうかはこれまでの結果からは分からないため切片を 2 から 5 まで変化させて推定を行い、同様に推定値の平均値のグラフを示す。

(図 7、 8)

切片が 2 の場合、データ生成過程で 1 が出される確率は 88.08% であり、切片が 5 の場合は 99.33% である[4]。図 7 と fig:切片と推定値 (beta)2 を見ると応答変数の分布が極端的に偏ると推定値が不安定になることが確認できる。今回の分析の場合、おおむね 92% までは推定値と母数がほぼ一致するが、それ以上になるとその差は急激に大きくなることが確認できる。

[4] x は-10 から 10 までの一様分布から抽出されたため、$E(x)=0$ である

1.4　パラメータを変える 2–説明変数

説明変数のパラメータを変えるということは第 1.3 節と同様に線形予測子を変えることを意味し、応答変数の分布に影響を与え、最終的には切片を変えることと同じような結果をもたらす。ただし、これは説明変数の分布によって異なる結果をもたらす。これまでの分析において x は $x \in (-10, 10)$ の一様分布から抽出されたため $E(x) = 0$ であり、説明変数のパラメータの変化は推定の安定性に影響を与えない。しかし、たとえば $x \in (0, 10)$ の場合、x は常に正の値をとるためパラメータの変化が応答変数の分布に影響を与え、結果的には推定の安定性まで影響を及ぼす。シミュレーションは割愛するが、説明変数 x のパラメータと x の分布による応答変数の理論的な分布の変化のみを表 4 で示す。

本節の分析を含めてこれまでの結果はあくまでも説明変数が一つのみのモデルのケースである。説明変数が追加されることによって必要とされるサンプルサイズも増加し (Long 1997)、また応答変数の分布と推定の安定性も変化しうる。

(表 4)

表 4 を見ると説明変数の期待値が 0 である場合、パラメータは応答変数に影響を与えないが、説明変数の期待値が 0 から遠く離れるとパラメータの変化に敏感になり従属変数の分布を大きく変えることが確認できる。

これまでの分析結果を総合して考えるとロジスティクス回帰分析の推定において重要なのはサンプルサイズ (n) と応答変数の分布である。小さいサンプルサイズと極端に偏る応答変数は不正確な推定を生み出すことが以上のシミュレーションと分析によって経験的に確認できる。

1.5　Omitted Variable

続いて、省略された変数による推定値のバイアスについて調べる。シミュレーションの手順はこれまでと同様であり、切片は 0.5、説明変数 $x1$ のパラメータは 0.2、省略された説明変数 $x2$ は 1 である。また、サンプルサイズは 500 であり、100 回施行し、切片と $x1$ のパラメータである β の推定値のプロットを図 9 と 10 で示す。

(図 9, 10)

結果を見ると省略された説明変数によって切片と β は母数から大きく離れたことが確認できる。これは OLS による線形回帰分析の結果と同様である[5]。

1.6 Overfitting

次は、省略された変数のシミュレーションと同様の条件で overfitting モデルをシミュレーションをする。データ生成過程において用いられるパラメータは $x1$ のみであるが、推定の際に $x2$ を投入し、図 11 から 14 で overfitting モデルと真のモデルの比較した。

(図 11 〜 14)

結果をみると overfitting モデルと新のモデルの間の推定値はほぼ同一であり、データ生成過程と全く関係のない説明変数によるバイアスは確認できない。しかし、OLS による線形回帰分析の結果と同様にモデルの効率性の問題がある。モデルの効率性の指標の一つである AIC で比較すると真のモデルが 533.5745、Overfitting モデルが 535.2508 であり、BIC を基準にすると真のモデルが 542.0037、Overfitting モデルは 547.8946 であることから Overfitting モデルは推定には大きな影響を与えないものの、効率性を損なう傾向があることが確認できる。

1.7 OLS との比較

(図 15、 16)

図 15 は OLS による線形確率モデル (Linear Probability Model; LPM) とロジスティクス回帰分析の予測確率を比較したグラフである。本課題において生成されたデータの場合、線形確率モデルとロジスティクスモデルの間には大きな違いが確認できない。2 つの予測確率の差は平均 0.007542、最小 0.00002705、最大 0.03225 であり、つまり予測確率は最大でも約 3 ポイントの差である。しかし、線形確率モデルの場合、図 16 のような分散不均一性という方法論的な問題以外にも実質的な解釈の面で問題がある。回帰分析の結果から新な説明変数を外挿して予測確率を算出するのは大きな意味はないが、理論的には線形確率モデルの場合、x が 16.59645 以上で $y \geq 1$、-17.89226 以下で $y \leq 0$ となり、確率としては無意味な値が得られるである。

[5] バイアスの大きさのメカニズムは異なる

2　問題 2

二値で表される変数と、その値を説明すると思われる変数（少なくとも 1 つは連続であることが望ましい）が含まれるデータを探しなさい。見つけたデータを CSV ファイルとして提出しなさい。(論文のレプリケーションデータも可)

(別途ファイルとして添付)

データセット名　CSES Module 3 (日本と韓国のみ抽出)
回答者数　2,373 名 (日本 1,373 名、韓国 1,000 名)
変数の数　445
調査時期　2007 年 (日本)、2008 年 (韓国)

3　シミュレーションのソースコード

```
logit.sim <- function(beta, n = 100, trials = 100, x.range =
    c(0, 10)){
 null.vector <- rep(NA, trials)

 data <- data.frame(
  b1 = null.vector,
  b1.se = null.vector,
  b1.ll = null.vector,
  b1.ul = null.vector,
  b2 = null.vector,
  b2.se = null.vector,
  b2.ll = null.vector,
  b2.ul = null.vector,
  para1 = null.vector,
  para2 = null.vector,
  wrap = null.vector)

 for(i in 1:trials){
  x <- runif(n, x.range[1], x.range[2])
  pr <- 1 / (1 + exp(-(beta[1] + beta[2] * x)))
  y <- rbinom(n = n, size = 1, prob = pr)

  fit <- glm(y ~ x, data = data, family=binomial(link="
      logit"))

  b1 <- coef(fit)[1]
  b2 <- coef(fit)[2]

  b1.se <- sqrt(summary(fit)$cov.unscaled[1,1])
  b2.se <- sqrt(summary(fit)$cov.unscaled[2,2])

  b1.ci95 <- confint(fit)[1,]
  b2.ci95 <- confint(fit)[2,]

  p1.temp <- b1.ci95[1] < beta[1] & b1.ci95[2] > beta[1]
  p2.temp <- b2.ci95[1] < beta[2] & b2.ci95[2] > beta[2]


  wrap <- rep(paste("n = ", n, ", trials = ", trials), trials)

  data[i,] <- c(b1,
                b1.se,
```

```
                b1.ci95,
                b2,
                b2.se,
                b2.ci95,
                p1.temp,
                p2.temp,
                wrap)
    print(paste("シミュレーション", i, "回目..."))

  }
  print("シミュレーション完了")
  return(data)
}
```

4 図表

表 1 ロジスティック回帰分析の結果

変数名	係数 (標準誤差)
intercept	0.25907^{**} (0.09745)
x	0.13978^{***} (0.01773)
N	500
$\chi^2_{df=499}$	690.26^{***}
$^{\dagger} \leq 0.1, ^{*} \leq 0.05, ^{**} \leq 0.01, ^{***} \leq 0.001$	

表 2 OLS の結果

変数名	係数 (標準誤差)
intercept	0.556019^{***} (0.020751)
x	0.031672^{***} (0.003579)
N	500
$F_{df=499}$	80.760^{***}
R^2	0.138
$^{\dagger} \leq 0.1, ^{*} \leq 0.05, ^{**} \leq 0.01, ^{***} \leq 0.001$	

表 3　切片の変化による応答変数の分布の変化

	0	0.1	0.5	1	5	10
頻度 *	49.4	52.2	60.7	70.27	98.7	100
最小値	37	40	50	60	97	100
最大値	63	68	71	81	99	100

*：DGP によって得られた 1 の数の平均

表 4　β と x の分布による応答変数の分布の変化

	$x \in (-10, 10)$	$x \in (0, 10)$	$x \in (-10, 0)$
$\beta = 0$	0.5000	0.5000	0.5000
$\beta = 0.1$	0.5000	0.6225	0.3775
$\beta = 0.5$	0.5000	0.9241	0.0759
$\beta = 1$	0.5000	0.9933	0.0067
$\beta = 5$	0.5000	1.0000	0.0000
$\beta = 10$	0.5000	1.0000	0.0000
	$x \in (-0.5, 0.5)$	$x \in (0, 0.5)$	$x \in (-0.5, 0)$
$\beta = 0$	0.5000	0.5000	0.5000
$\beta = 0.1$	0.5000	0.5062	0.4938
$\beta = 0.5$	0.5000	0.5312	0.4688
$\beta = 1$	0.5000	0.5622	0.4378
$\beta = 5$	0.5000	0.7773	0.2227
$\beta = 10$	0.5000	0.9241	0.0759

注：切片は 0 で固定

注 2：DGP で 1 が生成される確率

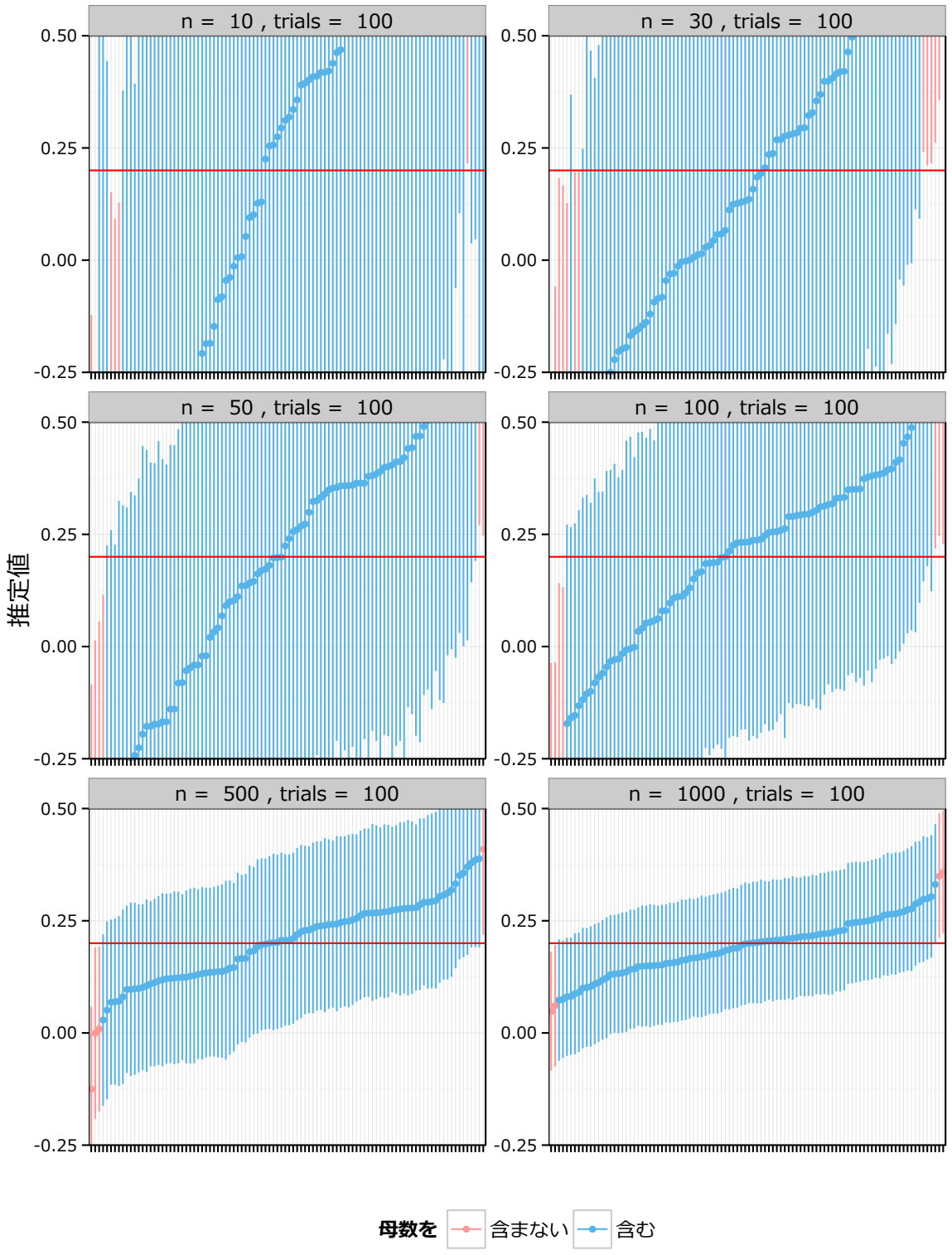

図 1　サンプルサイズと推定値 (切片)

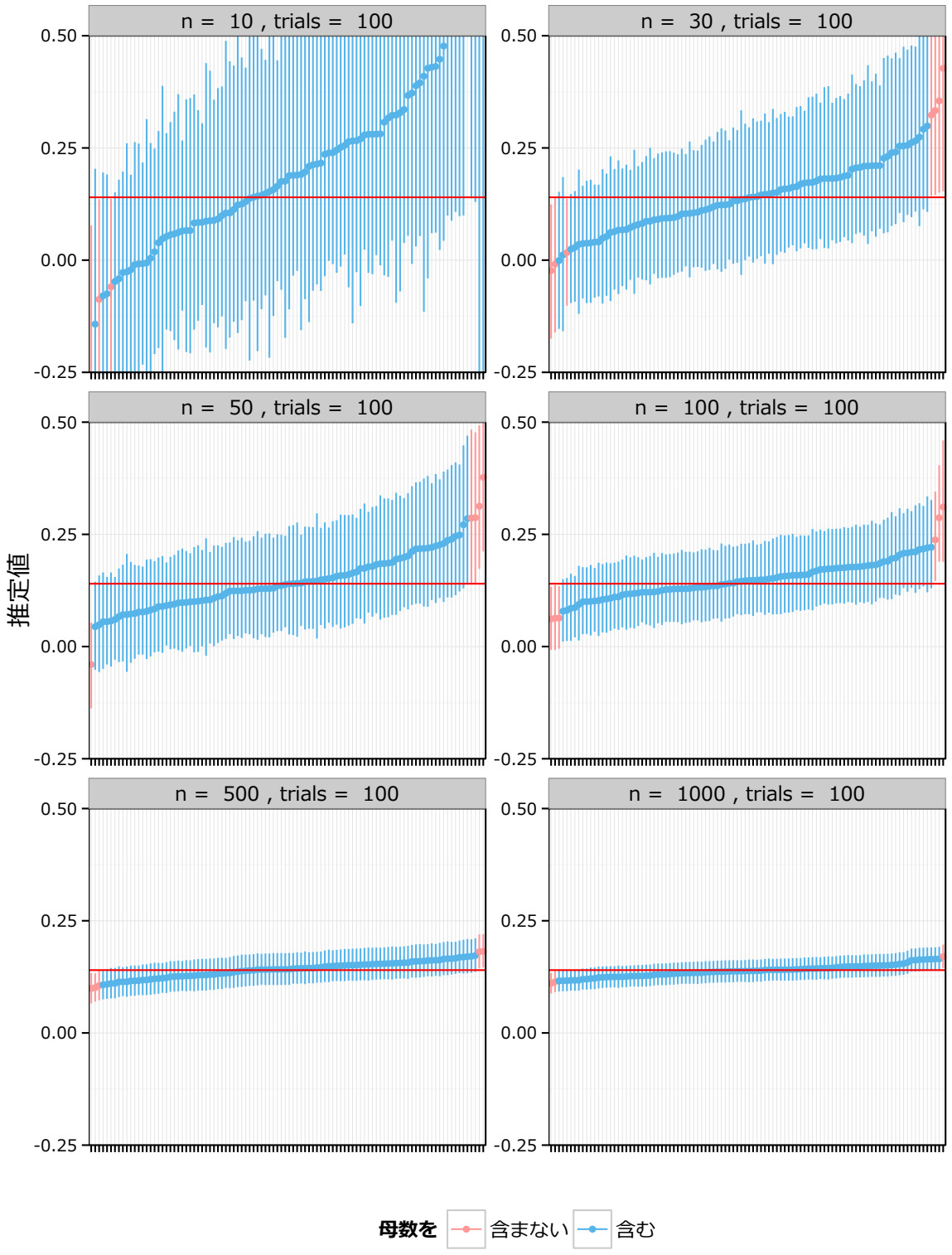

図 2　サンプルサイズと推定値 (β)

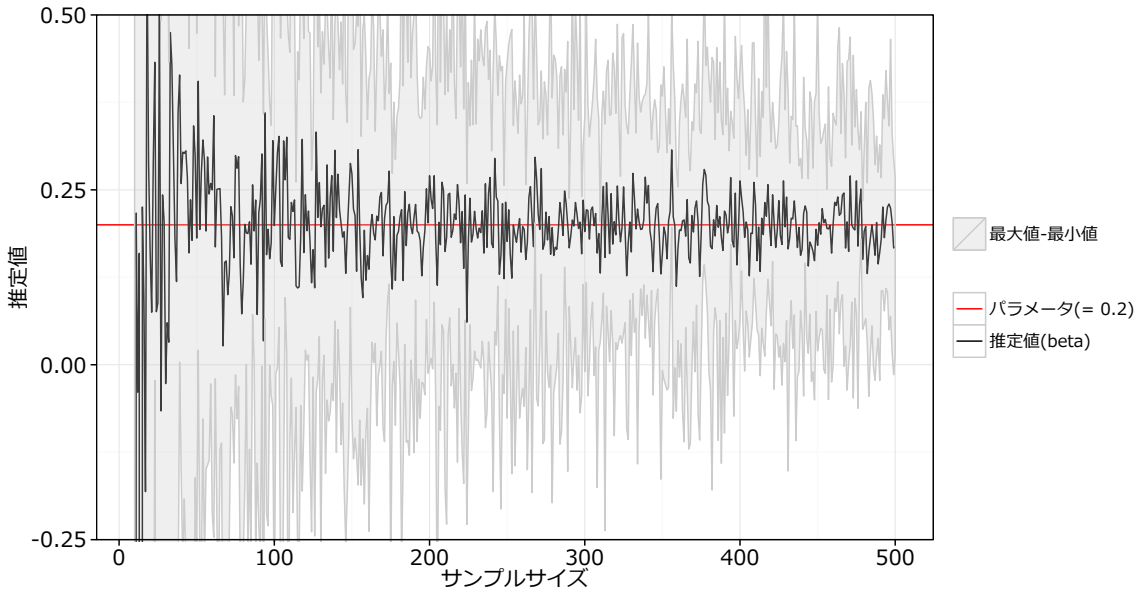

図 3　サンプルサイズと推定値の安定性 (切片)

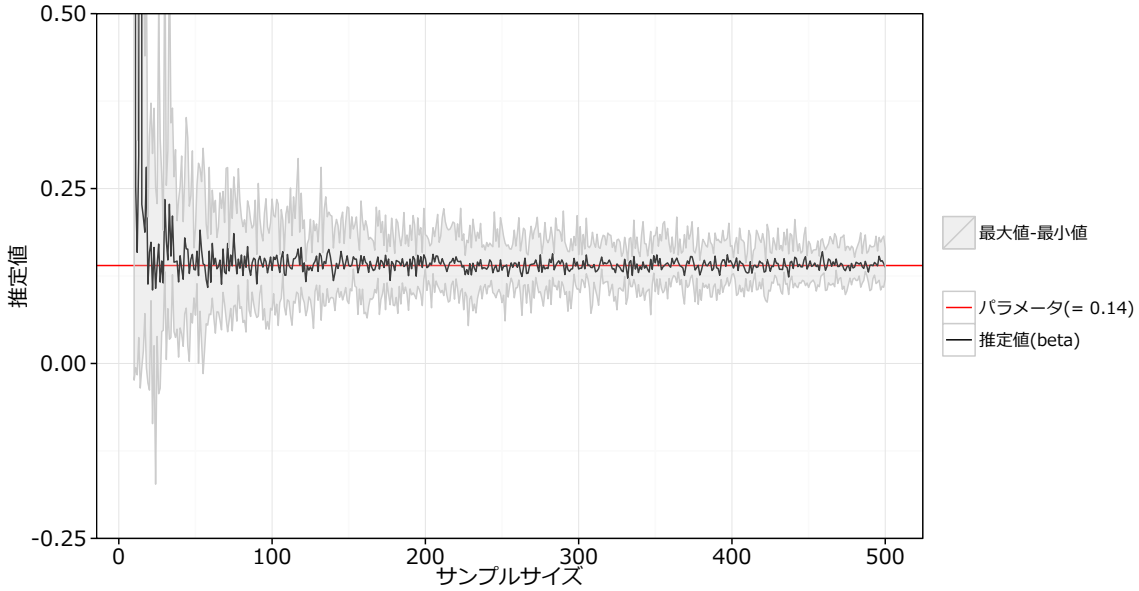

図 4　サンプルサイズと推定値の安定性 (beta)

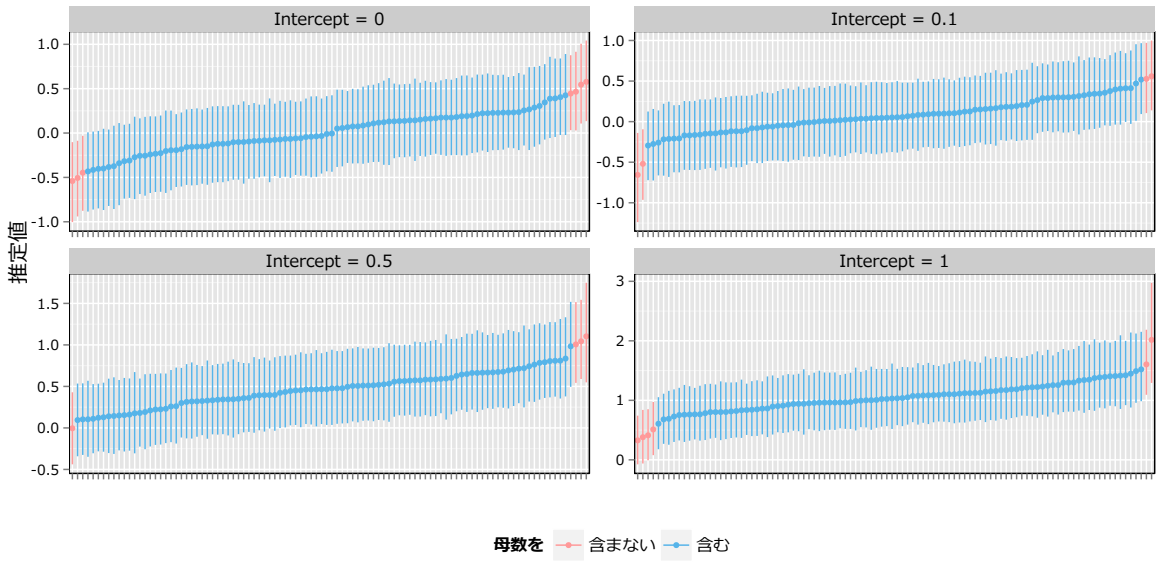

図 5　切片と推定値 (切片)

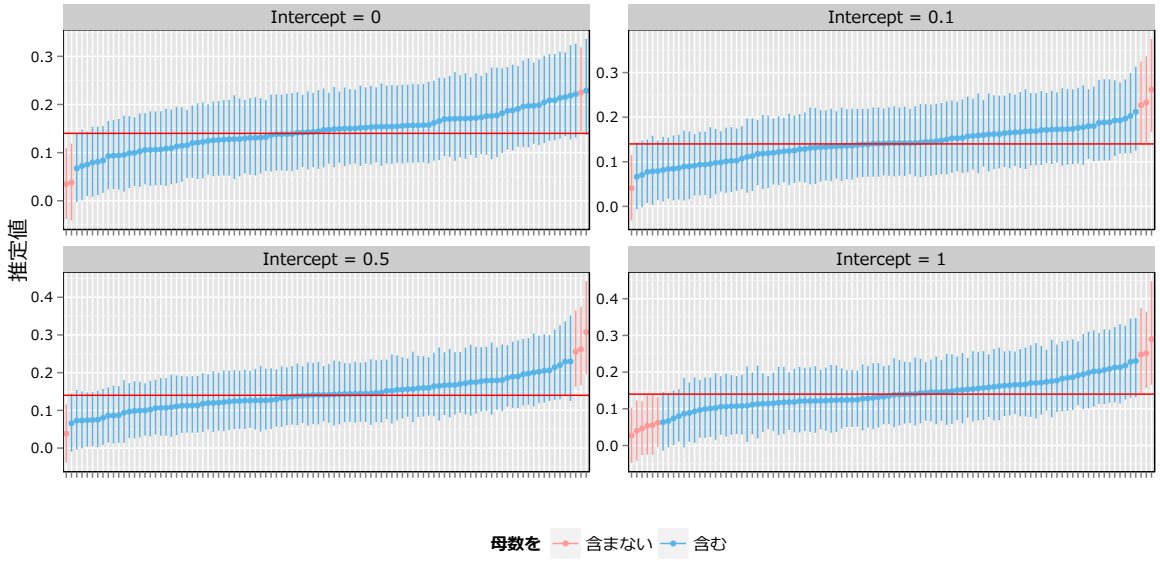

図 6　切片と推定値 (β)

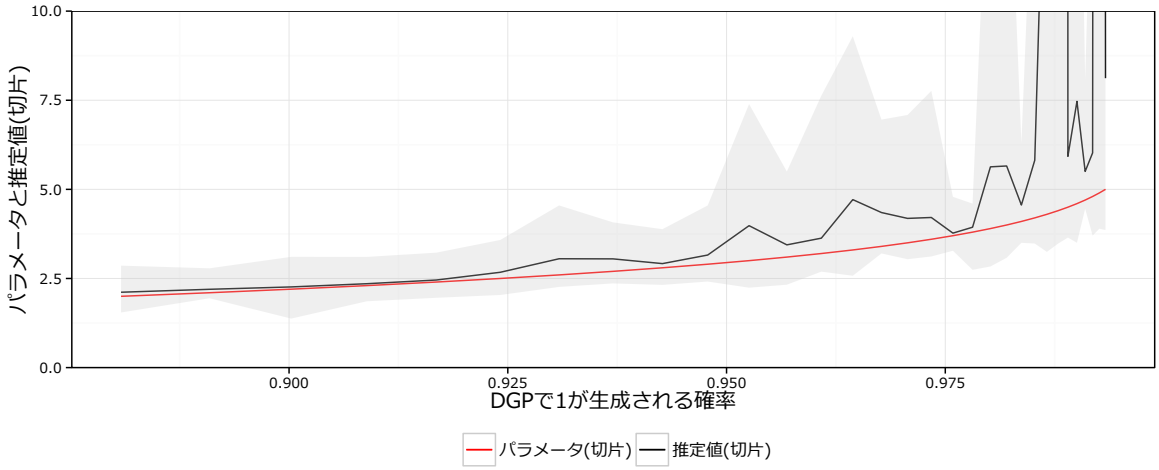

図 7　応答変数の分布が極端に偏る場合の切片と推定値 (切片)

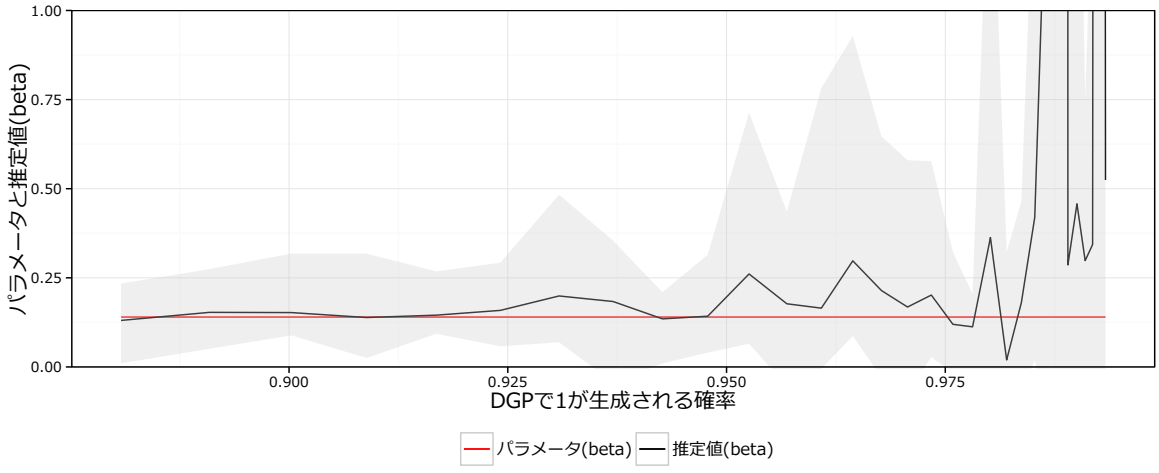

図 8　応答変数の分布が極端に偏る場合の切片と推定値 (β)

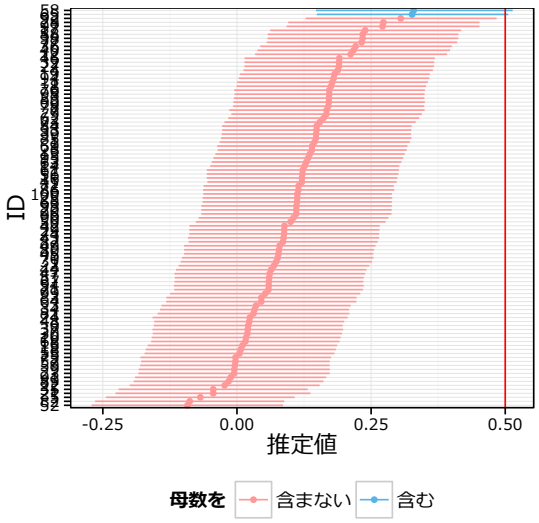

図 9　Omitted Variable Model(Intercept)

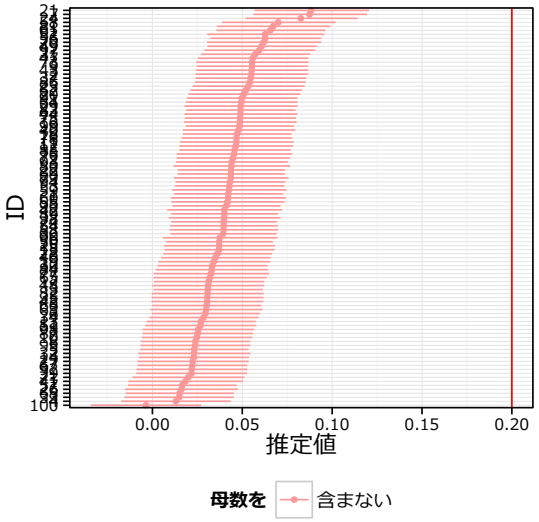

図 10　Omitted Variable Model(β)

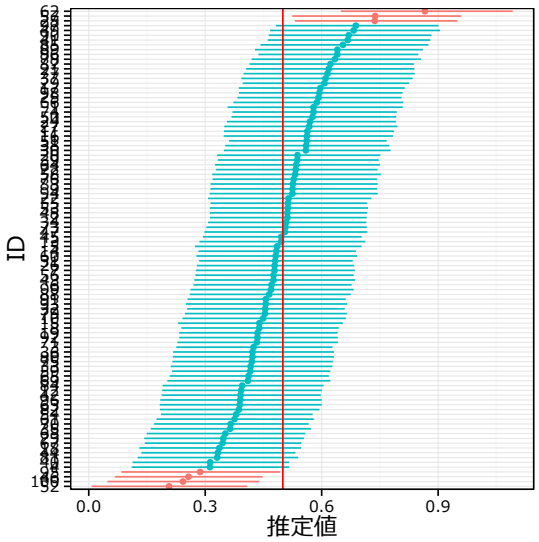

図 11　True Model(Intercept)

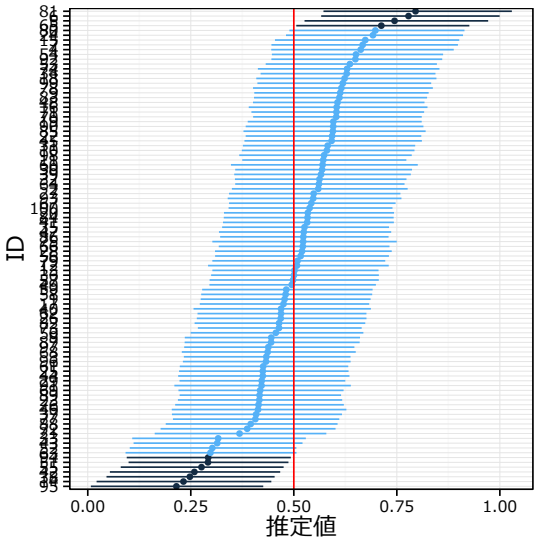

図 12　Overfitting Model(Intercept)

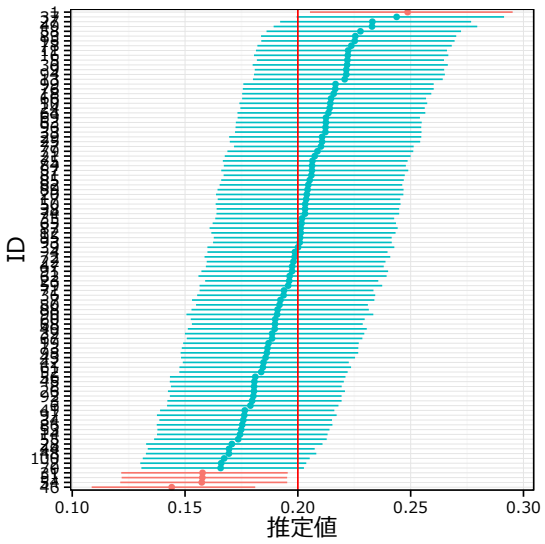

図 13　True Model(β)

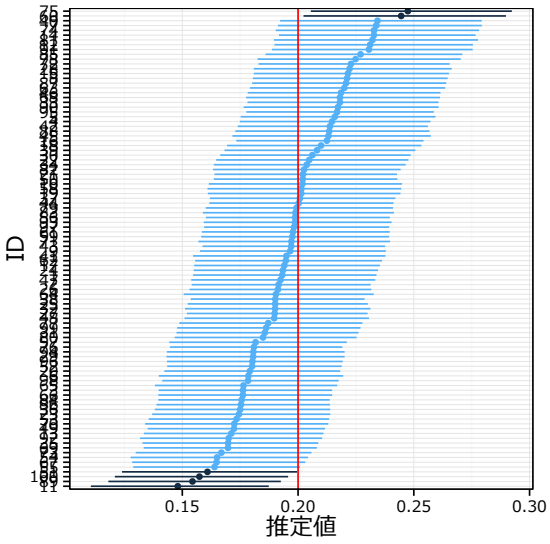

図 14　Overfitting Model(β)

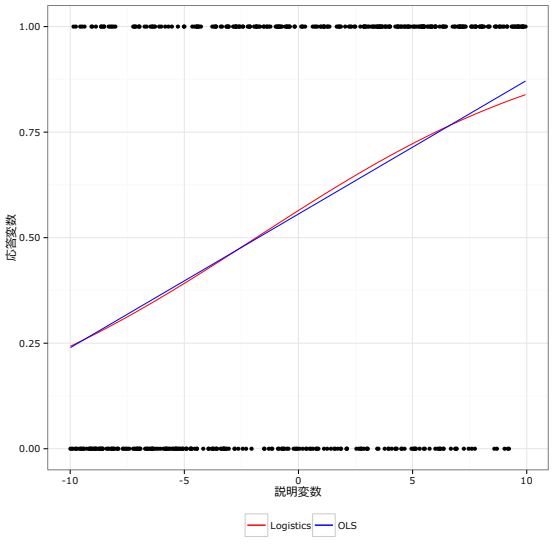

図 15　ロジスティックス回帰と OLS の比較

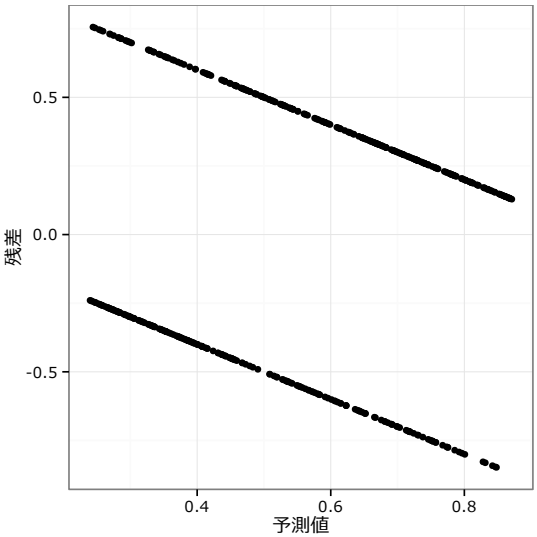

図 16　OLS の残差プロット