Ordinal response modelにおける ロバストダイバージェンスを用いた推定

東京理科大学大学院・創域理工学研究科　桃﨑 智隆
Tomotaka Momozaki
Department of Information Sciences, Tokyo University of Science

明星大学・データサイエンス学環　中川 智之
Tomoyuki Nakagawa
School of Data Science, Meisei University

1　導入

本稿は Momozaki and Nakagawa [1] の要約を与え, ordinal response model における外れ値の問題を取り扱う. 順序付きカテゴリカルデータは医学や社会科学などの幅広い分野で普及している. 例えば, 病気の進行度 (癌のステージ 1, 2, 3, 4) や, 政策に対する意見 (反対, 中立, 賛成) などがある. また, 年齢を 0〜20, 21〜40, 41〜60, 61〜80, 80 以上などのように, 連続データをカテゴリカルデータに要約した場合も順序付きカテゴリカルデータになる. このことから, 順序付きカテゴリカルデータは潜在的な連続変量を離散化したものと見做されることがある. 順序付きカテゴリカルデータの解析法は数多く研究されている [2]. 特に, 順序付きカテゴリカルデータを目的変数とし, それに対する回帰を考える ordinal response model は, 興味のある順序付きカテゴリカルデータと他のデータの関係を調べるのに重要であり [3], 例えば腫瘍学の分野では, 病気の進行度 (例えば, がんのステージ) において治療の傾向を判断するのに用いられている. Ordinal response model は, 順序付きカテゴリカルデータ y_i $(i = 1, 2, \ldots, n)$ の背景に連続潜在モデル $z_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ を仮定し, カットオフ $-\infty = \delta_0 < \delta_1 < \cdots < \delta_M = \infty$ を用いて y_i と z_i を $y_i = m \Leftrightarrow \delta_{m-1} < z_i \leq \delta_m$ で対応させたモデルである. ここで, n はサンプルサイズ, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ は共変量, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ は係数パラメータ, ε_i は確率密度関数 $g(\cdot)$ と分布関数 $G(\cdot)$ をもつ誤差項である. また, 観測されるデータは $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ と $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ であり, 潜在変数 $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^\top$ は観測されないものであり, 推定されるパラメータは $\boldsymbol{\beta}$ と $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_{M-1})^\top$ である. 分布関数 $G(\cdot)$ はリンク関数とも呼ばれ, 標準正規分布やロジスティック分布, left-skewed 対数ワイブル分布の分布関数がよく用いられ, それぞれプロビットリンク, ロジットリンク, 補対数対数リンクと呼ばれる.

Ordinal response model における $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top)^\top$ は一般的に最尤法を用いて推定されるが, 観測されるデータである $\boldsymbol{y}$ と $\boldsymbol{X}$ の組に外れ値が存在すると最尤推定量はその外れ値に強く影響を受けることが経験的に知られている. 外れ値は値の打ち間違いや単位の認識ミスなど, 様々な原因で発生する可能性がある. 連続データにおける外れ値に関する研

究は, 例えばHuber型ロス関数 [4] やロバストダイバージェンス [5, 6, 7] を用いたものなど, 数多くされているが, 離散型データである順序付きカテゴリカルデータを目的変数とするordinal response modelに対して簡単に議論を流用することはできない. 推論法が外れ値に対して頑健であるかを確認する方法は様々あるが, 一般的に影響関数が用いられる [8]. 影響関数とは, 簡単に述べると推論法におけるデータの影響度を確認するための指標である. あるデータに対してその値が発散してしまうと他のデータを無視して推論の結果がほとんどそのデータに依存してしまうので, 影響関数の値は有界であることが重要である. Scalera et al. [9] は, ordinal response modelにおける最尤法における影響関数が有界になるための条件を導出した. しかし, 一般的に使用されるプロビットやロジットなどのリンク関数はその条件を満たさない. そのため, 解析者は順序付きカテゴリカルデータに対してロバストかつ柔軟なモデリングを行うことができない. またリンク関数の誤特定はパラメータの推定に大きなバイアスをもたらす [10]. 影響関数の有界性は, 推論結果が特定のデータだけに影響を受けすぎないための重要な性質であるが, 大きく外れたデータが推論結果に全く影響を与えないというわけではない. そのため影響関数は有界性だけでなく, 再下降性 [11] を満たす, すなわち, 大きく外れたデータの影響が無視可能でもあることが望ましい.

本稿はordinal response modelにおいて, 最尤法における影響関数が密度関数をもつ分布のリンク関数に対して再下降性を満たさないことを示す. さらに我々は, 2つのロバストダイバージェンス, density-powerダイバージェンスとガンマダイバージェンスを用いたordinal response modelにおける推論法を提案し, 提案手法における影響関数が有界かつ再下降性を満たすためのリンク関数の条件を導出する. 一般的に使用されるリンク関数はこの条件を満たすため, 提案手法は外れ値に頑健かつリンク関数の選択に対して柔軟な推論を与える. 外れ値の割合と外れ値の大きさを変えた人工データを用いた数値実験を用いて, 平均二乗誤差や分類正答確率の観点から提案手法が最尤法よりも優れていることを示す. また, Affairsデータ [12] を用いて提案手法の頑健性と有用性を示す.

2 Ordinal response modelにおける最尤法

Ordinal response modelにおける尤度は以下となる.

$$f(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[G(\delta_{y_i} - \boldsymbol{x}_i^\top \boldsymbol{\beta}) - G(\delta_{y_i-1} - \boldsymbol{x}_i^\top \boldsymbol{\beta})\right]$$

この尤度において制約 $\delta_1 < \delta_2 < \cdots < \delta_{M-1}$ のもとでの最小化問題を解くことで $\boldsymbol{\theta}$ の最尤推定量が得られる. このような順序制約のもとでの最小化問題のためのアルゴリズムは多々あるが, 今回は $\boldsymbol{\delta}$ に対する再パラメータ化 $\delta_1 = \tilde{\delta}_1, \delta_m = \tilde{\delta}_1 + \sum_{j=2}^{m} \tilde{\delta}_j^2$ を用いる [13]. これにより, **R** における **stats** ライブラリの `optim` 関数などを用いることで容易に最小化問題を解くことができる.

推定手法における頑健性の指標として影響関数がよく使用される [8]. 今, 観測データ $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)$ は $F(y, \boldsymbol{x})$ から生成されたものとし, 汚染モデル $F_\rho(y, \boldsymbol{x}) = (1-\rho)F(y, \boldsymbol{x}) + \rho\Delta_{(y_o, \boldsymbol{x}_o)}(y, \boldsymbol{x})$ を考える. ただし, ρ は汚染割合, $\Delta_{(y_o, \boldsymbol{x}_o)}(y, \boldsymbol{x})$ は $(y_o, \boldsymbol{x}_o)$

で退化する汚染分布である. このとき最尤法における影響関数は以下となる.

$$IF_{ML}(y_o, \boldsymbol{x}_o; F, \boldsymbol{\theta}) = n^{-1}\mathcal{I}^{-1}(\boldsymbol{\theta})\frac{\partial \log f(y_o|\boldsymbol{x}_o;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

ただし, $\mathcal{I}(\boldsymbol{\theta})$ はフィッシャー情報行列である. Scalera et al. [9] は, ordinal response model において最尤法における影響関数が有界になるための以下の条件を導出した.

定理 1 (Scalera et al. [9] の補題 3.1 と 3.2) *Ordinal response model* において, 最尤法における $\boldsymbol{\beta}$ と $\boldsymbol{\delta}$ の影響関数が有界になるための必要十分条件はそれぞれ以下である.

$$\lim_{u\to\pm\infty}\left|u\frac{\partial \log g(u)}{\partial u}\right| < +\infty, \quad \lim_{u\to\pm\infty}\left|\frac{\partial \log g(u)}{\partial u}\right| < +\infty$$

これらの条件を満たすリンク関数として, 自由度が十分に小さいスチューデントの t 分布の分布関数が挙げられるが, プロビットリンクなどの一般的に使用されるリンク関数はこれらを満たさない [9]. Scalera et al. [9] において, 影響関数の有界性に関する議論はされているが, 再下降性に関する議論はされていない. 我々はこの影響関数における再下降性について以下の定理を導出した.

定理 2 *Ordinal response model* において, 最尤法におけるパラメータ $\boldsymbol{\beta}$ と $\boldsymbol{\delta}$ の影響関数が再下降性を満たすためには, $x \to \infty$ でリンク関数の一階導関数が $(\log x)^{-1}$ のオーダーをもつ必要がある

証明は Momozaki and Nakagawa [1] の Appendix を参照されたい. この定理から ordinal response model において, 最尤法における影響関数が密度関数をもつ分布のリンク関数に対して再下降性を満たさなく, 大きく外れたデータの影響を完全に除去できないことがわかる.

3　ダイバージェンスを用いた提案ロバスト推定法

ダイバージェンスとして density-power ダイバージェンス [5] とガンマダイバージェンス [6, 7] を用いた, ordinal response model におけるパラメータ $\boldsymbol{\theta}$ の以下のロバスト推定量を提案する.

$$\hat{\boldsymbol{\theta}}_{DP} = \arg\min_{\boldsymbol{\theta}} \tilde{d}_{DP}(f(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta})), \quad \hat{\boldsymbol{\theta}}_{\gamma} = \arg\min_{\boldsymbol{\theta}} \tilde{d}_{\gamma}(f(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta}))$$

ここで, $\tilde{d}_{DP}(f(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta}))$ と $\tilde{d}_{\gamma}(f(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta}))$ は density-power ダイバージェンスとガンマダイバージェンスにおける相互エントロピーの経験推定値であり, チューニングパラメータ $\alpha, \gamma > 0$ に対してそれぞれ以下で表される.

$$\begin{aligned}\tilde{d}_{DP}(f(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta})) = &-\frac{1}{\alpha}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[G(\delta_{y_i} - \boldsymbol{x}_i^\top\boldsymbol{\beta}) - G(\delta_{y_i-1} - \boldsymbol{x}_i^\top\boldsymbol{\beta})\right]^{\alpha}\right\} \\ &+ \frac{1}{1+\alpha}\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}\left[G(\delta_m - \boldsymbol{x}_i^\top\boldsymbol{\beta}) - G(\delta_{m-1} - \boldsymbol{x}_i^\top\boldsymbol{\beta})\right]^{1+\alpha}\right\},\end{aligned}$$

$$\tilde{d}_\gamma(f(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta})) = -\frac{1}{\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\left[G(\delta_{y_i}-\boldsymbol{x}_i^\top\boldsymbol{\beta})-G(\delta_{y_i-1}-\boldsymbol{x}_i^\top\boldsymbol{\beta})\right]^\gamma\right\}$$
$$+\frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}\left[G(\delta_m-\boldsymbol{x}_i^\top\boldsymbol{\beta})-G(\delta_{m-1}-\boldsymbol{x}_i^\top\boldsymbol{\beta})\right]^{1+\gamma}\right\}$$

また, density-power ダイバージェンスとガンマダイバージェンスを用いた ordinal response model における影響関数はそれぞれ以下で表される.

$$\begin{aligned}
&IF_{DP}(y_o,\boldsymbol{x}_o;F,\beta_k)\\
\propto&-\left[g(\delta_{y_o}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-g(\delta_{y_o-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]\left[G(\delta_{y_o}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{y_o-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\alpha-1}x_{ok}\\
&+\left(\sum_{m=1}^{M}\left[g(\delta_m-\boldsymbol{x}_o^\top\boldsymbol{\beta})-g(\delta_{m-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]\left[G(\delta_m-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{m-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\alpha}\right)x_{ok},
\end{aligned}$$

$$\begin{aligned}
&IF_{DP}(y_o,\boldsymbol{x}_o;F,\delta_l)\\
\propto&g(\delta_l-\boldsymbol{x}_o^\top\boldsymbol{\beta})\left[G(\delta_{y_o}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{y_o-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\alpha-1}\left[I(y_o=l)-I(y_o=l+1)\right]\\
&-g(\delta_l-\boldsymbol{x}_o^\top\boldsymbol{\beta})\left(\left[G(\delta_l-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{l-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\alpha}-\left[G(\delta_{l+1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_l-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\alpha}\right),
\end{aligned}$$

$$\begin{aligned}
&IF_{\gamma}(y_o,\boldsymbol{x}_o;F,\beta_k)\\
\propto&-\left(\sum_{m=1}^{M}\left[G(\delta_m-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{m-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma+1}\right)\\
&\quad\times\left[g(\delta_{y_o}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-g(\delta_{y_o-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]\left[G(\delta_{y_o}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{y_o-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma-1}x_{ok}\\
&+\left[G(\delta_{y_o}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{y_o-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma}x_{ok}\\
&\quad\times\left(\sum_{m=1}^{M}\left[g(\delta_m-\boldsymbol{x}_o^\top\boldsymbol{\beta})-g(\delta_{m-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]\left[G(\delta_m-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{m-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma}\right),
\end{aligned}$$

$$\begin{aligned}
&IF_{\gamma}(y_o,\boldsymbol{x}_o;F,\delta_l)\\
\propto&\left(\sum_{m=1}^{M}\left[G(\delta_m-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{m-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma+1}\right)\\
&\quad\times g(\delta_l-\boldsymbol{x}_o^\top\boldsymbol{\beta})\left[G(\delta_{y_o}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{y_o-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma-1}\left[I(y_o=l)-I(y_o=l+1)\right]\\
&-g(\delta_l-\boldsymbol{x}_o^\top\boldsymbol{\beta})\left[G(\delta_{y_o}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{y_o-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma}\\
&\quad\times\left(\left[G(\delta_l-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_{l-1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma}-\left[G(\delta_{l+1}-\boldsymbol{x}_o^\top\boldsymbol{\beta})-G(\delta_l-\boldsymbol{x}_o^\top\boldsymbol{\beta})\right]^{\gamma}\right)
\end{aligned}$$

これら影響関数の有界性と再下降性について以下の定理を導出した.

定理 3

$$\lim_{u\to\pm\infty}g(u)^\alpha u=0 \tag{1}$$

を満たすような $0<\alpha\leq 1$ $(0<\gamma\leq 1)$ が存在するならば, *density-power (ガンマ) ダイバージェンスを用いた ordinal response model* における影響関数は有界性と再下降性を満たす.

証明は Momozaki and Nakagawa [1] の Appendix を参照されたい. 一般的に使用されるリンク関数であるプロビットやロジット, 補対数対数リンクにおいて, 条件式 (1) を満たすような $0 < \alpha \leq 1$ $(0 < \gamma \leq 1)$ が存在する. よって, 解析者は外れ値に頑健かつ, リンク関数を選択できるため自由なモデリングを行うことができる.

図 1 は, $p = 1$, $\beta = 1.0$, $\boldsymbol{\delta} = (-1.5, 0, 1.5)$ とし, プロビット (上段) とロジット (下段) リンクを用いた際の最尤法における影響関数 (黒色) と, density-power ダイバージェンス (青色) とガンマダイバージェンス (赤色) を用いた際の影響関数のプロットである. また, 左が係数パラメータ β, 右がカットポイント δ_1 の影響関数を表しており, 横軸は共変量の値, 縦軸は影響関数の値である. この図からわかるように, 最尤法における影響関数の値は共変量の値に対して発散, もしくはゼロ以外のある値に収束しているため再下降性を満たしていないことがわかる. 一方で, 提案手法における影響関数の値は共変量の値に対してゼロに収束しており, プロビットとロジットリンクを用いた際の影響関数は有界性と再下降性を満たしていることがわかる.

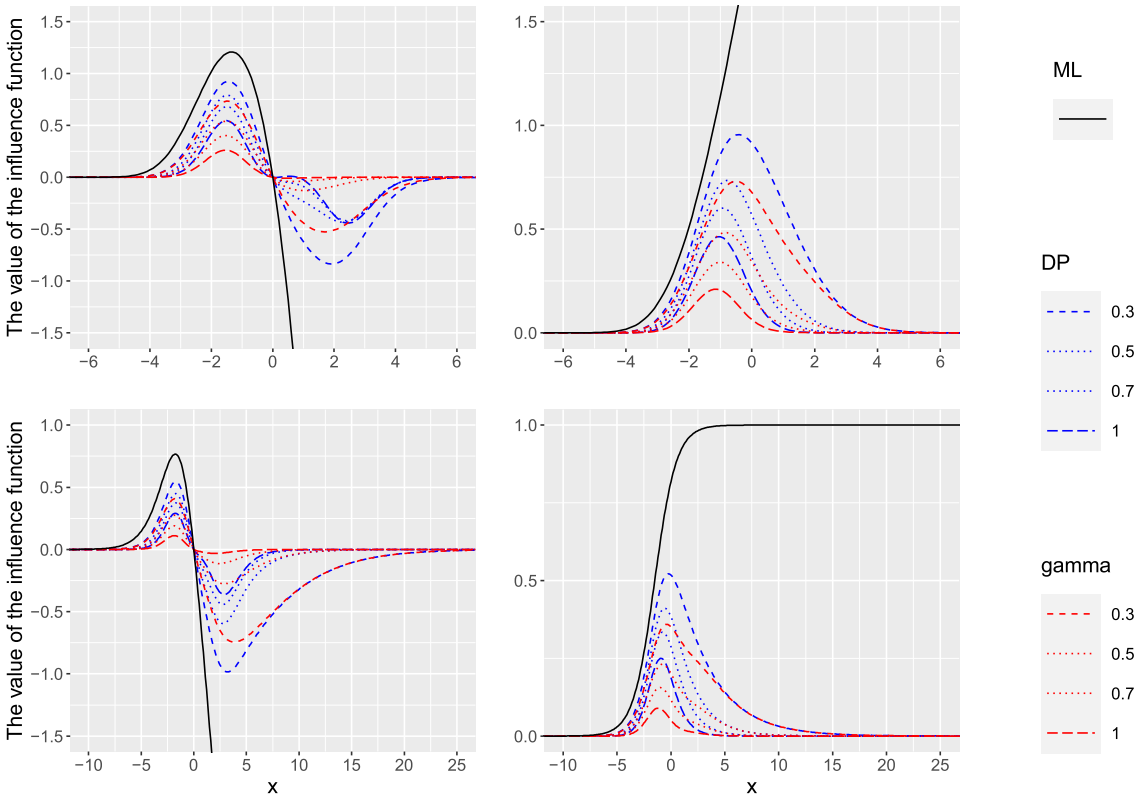


図 1: プロビット (上段) とロジットリンク (下段) を用いた際の最尤法における影響関数 (黒色) と, density-power ダイバージェンス (青色) とガンマダイバージェンス (赤色) を用いた際の影響関数のプロット.

4 数値実験

5 つのカテゴリをもつ順序付きカテゴリカルデータに対して, 連続潜在モデル $z_i = x_i\beta_1 + d_i\beta_2 + x_i d_i \beta_3 + \varepsilon_i$ を考える. 係数パラメータは $(\beta_1, \beta_2, \beta_3) = (2.5, 1.2, 0.7)$ とし, 誤差項

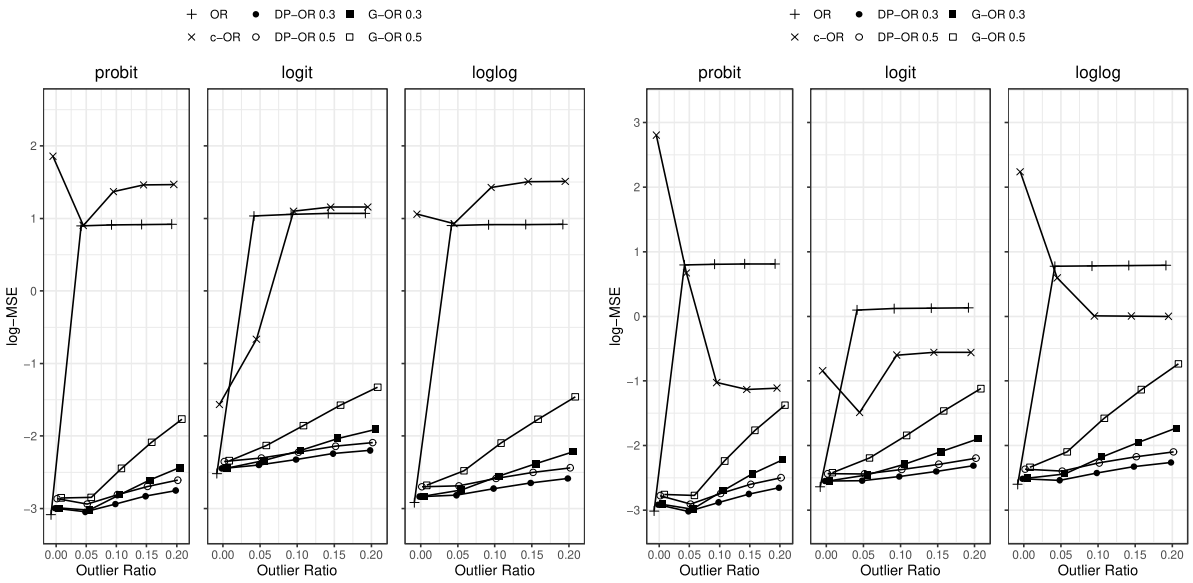
ε_i は標準正規 (probit), ロジスティック (logit), ガンベル (log-log) 分布に従うとし, そのときのカットオフはぞれぞれ $(\delta_1, \delta_2, \delta_3, \delta_4) = (-3.0, -0.7, 1.6, 3.9)$, $(-3.3, -0.8, 1.7, 4.2)$, $(-2.9, 1.0, 2.9, 4.8)$ とした. また, $x_i \sim (1-\rho)N(0,1) + \rho N(\mu, 1)$, $d_i \sim$ Bernoulli(0.25) であり, x_i と d_i は互いに独立とする. 上記の設定で, 汚染分布の平均 μ を 20 と固定し外れ値の割合を $\rho = (0, 0.05, 0.10, 0.15, 0.20)$ とした場合と, $\rho = 0.05$ と固定し $\mu = (0, 5, 10, 15, 20)$ とした場合のもとで, サンプル数 $n = 200$ のデータ生成を $S = 10000$ 回行った. 評価指標は平均二乗誤差 (MSE) $p^{-1}\sum_{j=1}^{p}(\hat{\theta}_j - \theta_j)^2$ と分類正答確率 (CCR) $S^{-1}\sum_{s=1}^{S} I(\hat{y}_s = y_s^*)$ を用いた. ただし, $\hat{y}_s$ は予測値, y_s^* はテストデータ, $I(\cdot)$ は定義関数を表す. 提案手法と比較する手法として, 誤差分布に対応するリンク関数を用いた最尤法の他に, 各誤差分布よりも裾の重いコーシー分布の分布関数である cauchit リンクを用いた最尤法を使用した.

図 2 と 3 はそれぞれ外れ値の割合 ρ と汚染分布の平均 μ の変化に対して, 各手法における対数 MSE の値をプロットしたものである. 十字 (+) とばつ印 (×) は, 誤差分布に対応するリンク関数を用いた最尤法と cauchit リンクを用いた最尤法を表す (OR, c-OR). また, density-power ダイバージェンスを用いた提案手法とガンマダイバージェンスを用いた提案手法をぞれぞれ丸印 (○) と四角印 (□) で表し (DP-OR, G-OR), チューニングパラメータの値に 0.3 を使用したものを黒, 0.5 を使用したものを白とした外れ値がない場合には, 最尤法と提案手法は同程度の精度を示している. 一方で, 外れ値がある場合には, 最尤法は推定精度がかなり悪くなってしまうのに対して, 提案手法は比較的安定した推定精度を保つことができている. 表 1 は, ρ と μ の変化に対する各手法における CCR の値を示している. 対数 MSE の結果と同様に, 外れ値がない場合には最尤法と提案手法は同程度の精度を示しているが, 外れ値がある場合に最尤法は CCR の値が低い, 特に μ の値が大きくなるにつれて CCR の値が低くなっているのに対して提案手法は ρ や μ の値が大きくなってもほとんど変わらない結果を示している. 以上の結果から, 提案手法は外れ値の割合 ρ や汚染分布の平均 μ の変化に対して頑健な推定精度をもつことがわかった.

表 1: 外れ値の割合 ρ と汚染分布の平均 μ の変化に対する各手法の CCR の値

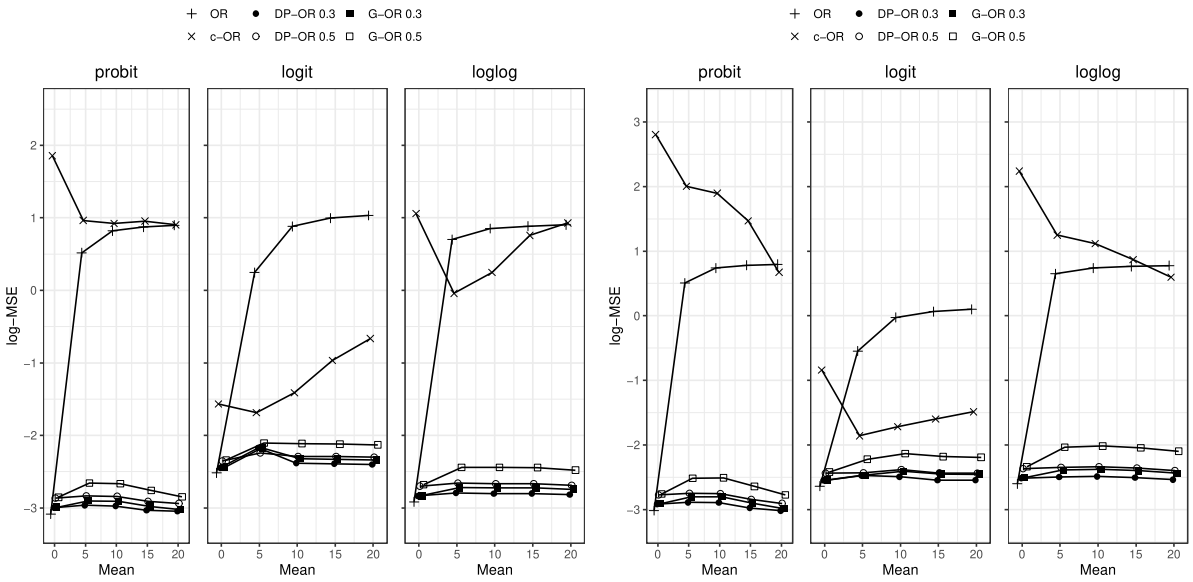
		OR	c-OR	DP-OR		G-OR	
link	ρ			0.3	0.5	0.3	0.5
probit	0%	0.6854	0.6827	0.6859	0.6862	0.6859	0.6862
	5%	0.4162	0.5779	0.6829	0.6834	0.6832	0.6830
	10%	0.4211	0.4341	0.6805	0.6794	0.6803	0.6795
	15%	0.4211	0.4220	0.6806	0.6781	0.6806	0.6781
	20%	0.4162	0.4140	0.6766	0.6764	0.6760	0.6764
logit	0%	0.5441	0.5397	0.5450	0.5455	0.5451	0.5455
	5%	0.3819	0.4696	0.5417	0.5401	0.5416	0.5412
	10%	0.3745	0.3735	0.5370	0.5376	0.5368	0.5375
	15%	0.3711	0.3707	0.5317	0.5317	0.5321	0.5319
	20%	0.3750	0.3725	0.5361	0.5364	0.5367	0.5357
loglog	0%	0.6638	0.6614	0.6632	0.6637	0.6632	0.6632
	5%	0.5044	0.5742	0.6684	0.6682	0.6681	0.6693
	10%	0.5092	0.5086	0.6653	0.6637	0.6651	0.6633
	15%	0.4994	0.4934	0.6659	0.6666	0.6660	0.6670
	20%	0.4991	0.4916	0.6640	0.6643	0.6645	0.6626

	OR	c-OR	DP-OR		G-OR	
μ			0.3	0.5	0.3	0.5
0	0.6854	0.6827	0.6859	0.6862	0.6859	0.6862
5	0.5117	0.6722	0.6778	0.6770	0.6780	0.6771
10	0.4304	0.6672	0.6828	0.6839	0.6826	0.6828
15	0.4227	0.6327	0.6800	0.6788	0.6802	0.6788
20	0.4162	0.5779	0.6829	0.6834	0.6832	0.6830
0	0.5441	0.5397	0.5450	0.5455	0.5451	0.5455
5	0.4579	0.5189	0.5329	0.5385	0.5337	0.5369
10	0.3933	0.5061	0.5337	0.5337	0.5330	0.5337
15	0.3816	0.4841	0.5408	0.5403	0.5402	0.5409
20	0.3819	0.4696	0.5417	0.5401	0.5416	0.5412
0	0.6638	0.6614	0.6632	0.6637	0.6632	0.6632
5	0.5015	0.6733	0.6818	0.6813	0.6814	0.6815
10	0.5142	0.6454	0.6660	0.6657	0.6657	0.6656
15	0.5058	0.6014	0.6764	0.6758	0.6769	0.6752
20	0.5044	0.5742	0.6684	0.6682	0.6681	0.6693



(a) 係数パラメータ $\boldsymbol{\beta}$ (b) カットオフ $\boldsymbol{\delta}$

図 2: 外れ値の割合 ρ の変化に対する各手法の対数 MSE のプロット



(a) 係数パラメータ $\boldsymbol{\beta}$ (b) カットオフ $\boldsymbol{\delta}$

図 3: 汚染分布の平均 μ の変化に対する各手法の対数 MSE のプロット

5　実データ解析

サンプル数 601 と 9 つの変数 (3 つの連続変数, 2 つの二値変数, 2 つの順序付きカテゴリカル変数, 2 つの名義カテゴリカル変数) を含む Affairs データ [12] を考える. ここでは, 順序付きカテゴリカル変数 “self-rating of marriage” を目的変数とし, 残りの変数を共変量とし, 連続変数は平均 0, 分散 1 になるように標準化, 順序付きカテゴリカル変数はリッカート・シグマ法を用いて変換, 名義カテゴリカル変数はダミー変数化した. また, リンク関数には一般的に使用されるプロビットとロジットリンクを用いた.

図 4 は, Affairs データについて 2 つのリンク関数を用いた一般化残差 [13] の値をプロットしたものである. 実線と破線はそれぞれ一般化残差の値の 95%区間と 99%区間を示す. 一般化残差プロットは, データにおける外れ値の存在を確認するためによく使用される [13]. 図 4 の一般化残差の値が 95%区間 (実線) より大きい標本を「外れ値」と仮定する. 総サンプルの 8 割を訓練データとし, 残りの 2 割を検証データとした. ただし, 検証データには外れ値が含まれないようにした. 訓練データから外れ値を除いたものを「修正データ」, 除いていないものを「元データ」とする. これらを用いて, 最尤法を用いた手法 (OR) と density-power ダイバージェンスを用いた提案手法 (DP-OR) とガンマダイバージェンスを用いた提案手法 (G-OR) を比較する. ここで, チューニングパラメータの値には 0.3 と 0.5 を用い, 評価指標に $q^{-1}\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^c\|_2^2$ と元データを用いた際の分類正答率を用いた. ただし, $\hat{\boldsymbol{\theta}}$ は元データを用いた際の推定値, $\hat{\boldsymbol{\theta}}^c$ は修正データを用いた際の推定値, q は総パラメータ数を表す.

表 2 はプロビットとロジットリンクを用いた際の $q^{-1}\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^c\|_2^2$ の値を示しており, 最尤法よりも提案手法の方が値が小さくなっている. このことから, パラメータの推定において, 最尤法と比較して提案手法の方が外れ値の影響を受けていないことがわかる. また表 3 はプロビットとロジットリンクを用いた際の分類正答率を示しており, 最尤法と比較して提案手法の方が値が大きくなっており, 検証データにおける目的変数 “self-rating of marriage” を良く予測できていることがわかる.

表 2: $q^{-1}\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^c\|_2^2$ の値

(a) プロビットリンク

	OR	DP-OR		G-OR	
		0.3	0.5	0.3	0.5
$\boldsymbol{\beta}$	0.5549	0.0681	0.0413	0.0725	0.0623
$\boldsymbol{\delta}$	1.0236	0.1351	0.0854	0.1399	0.1082

(b) ロジットリンク

	OR	DP-OR		G-OR	
		0.3	0.5	0.3	0.5
$\boldsymbol{\beta}$	1.3836	0.1401	0.0798	0.1277	0.1197
$\boldsymbol{\delta}$	2.6987	0.3594	0.2190	0.3474	0.2536

表 3: 分類正答率

(a) プロビットリンク

OR	DP-OR		G-OR	
	0.3	0.5	0.3	0.5
0.4793	0.5041	0.4959	0.5041	0.5041

(b) ロジットリンク

OR	DP-OR		G-OR	
	0.3	0.5	0.3	0.5
0.4793	0.4959	0.4959	0.4959	0.4959

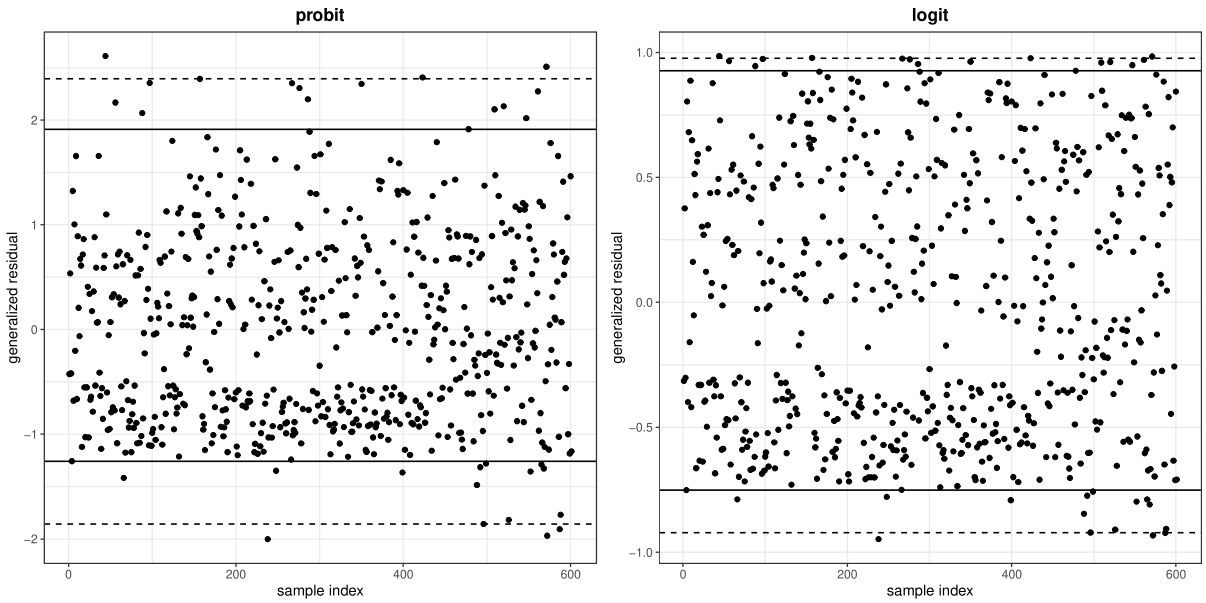


図 4: プロビットとロジットリンクによる Affairs データにおける一般化残差 [13] の値のプロット, 実線と破線はそれぞれ一般化残差の値の 95%区間と 99%区間を示す.

6 まとめ

本稿は Momozaki and Nakagawa [1] の要約を与え, ordinal response model における外れ値の問題を取り扱った. Ordinal response model において, 最尤法における影響関数が密度関数をもつ分布のリンク関数に対して再下降性を満たさない, すなわち, 大きく外れたデータの影響を無視することができないことを示した. また, density-power ダイバージェンスとガンマダイバージェンスを用いた ordinal response model におけるロバストな推定法を提案した. 提案手法における影響関数も導出し, それらが有界性と再下降性を満たすための条件を導出した. 一般的に使用されるリンク関数であるプロビットやロジット, 補対数対数リンクはこの条件を満たす. 数値実験では, 外れ値の割合と外れ具合を変えた人工データを用いて, バイアスや平均二乗誤差, 分類正答確率の観点から提案手法と最尤法を比較した. 結果, 提案手法は外れ値が存在しない場合には最尤法と同程度の推定精度を示し, また, 外れ値の割合や外れ具合が大きくなっても提案手法はそれらに影響を受けない推定精度を示した. さらに, 実データ解析例として Affairs データを取り上げ, 最尤法よりも提案手法の方が外れ値の影響を受けにくく予測精度が高いことを示した.

謝辞

RIMS 共同研究「種々の統計的モデルにおける推測方式の有効性」にて貴重な講演の機会を頂き, 有難うございます. また本研究を行うにあたり, 有益なコメントを頂いた慶應義塾大学の菅澤翔之助先生と東京大学の入江薫先生に感謝申し上げます. 最後に, 本研究は JSPS 科研費 19K14597 と 21H00699 の助成を受けたものです.

参考文献

[1] Momozaki, T. and Nakagawa, T. (2022). Robustness against outliers in ordinal response model via divergence approach. *arXiv preprint arXiv:2209.11965*

[2] Agresti, A. (2010). *Analysis of Ordinal Categorical Data.* John Wiley & Sons.

[3] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* **42**, 109–127.

[4] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics.* Wiley Online Library

[5] Basu, A., Harris, I. R., Hjort, N. L., and Jones, MC. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–559.

[6] Jones, MC., Hjort, N. L., Harris, I. R., and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88**, 865–873.

[7] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis* **99**, 2053–2081.

[8] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

[9] Scalera, V., Iannario, M, and Monti, A. C. (2021). Robust link functions. *Statistics* **55**, 963–977.

[10] Czado, C. and Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference* **33**, 213–231.

[11] Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R).* John Wiley & Sons.

[12] Fair, R. C. (1978). A theory of extramarital affairs. *Journal of Political Economy* **86**, 45–61.

[13] Franses, P. H. and Paap, R. (2001). *Quantitative Models in Marketing Research.* Cambridge University Press.

連絡先　桃﨑 智隆　tomotaka.8823.momozaki[あっと]gmail.com　TEL: 04-7124-1501

〒 278-8510 千葉県野田市山崎 2641

東京理科大学大学院 創域理工学研究科 情報計算科学専攻