Reduced-bias estimation for models with ordinal responses

Ioannis Kosmidis
i.kosmidis@ucl.ac.uk
http://ucl.ac.uk/~ucakiko

Department of Statistical Science, University College London
The Alan Turing Institute

31 August 2017
CEN ISBS 2017 Joint Conference
Vienna, Austria
Outline

1. Testing for proportional odds
2. Reducing bias
3. Direction of shrinkage
4. Discussion
Outline

1. Testing for proportional odds
2. Reducing bias
3. Direction of shrinkage
4. Discussion
Wine tasting data\(^1\)

<table>
<thead>
<tr>
<th>contact</th>
<th>temp</th>
<th>rating</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>no</td>
<td>cold</td>
<td>4</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>no</td>
<td>warm</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>yes</td>
<td>cold</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>yes</td>
<td>warm</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

Experiment on the effect of factors on the bitterness of white wine

c**ontact** of juice with skin and **temp**erature when crushing the grapes
9 judges rated 2 bottles per combination of factors in terms of bitterness

\(^1\)data from Randall (1989)
Empirical cumulative logits for factor combination $i$ and rating $j$

$$\log \frac{Y_{i1} + \ldots + Y_{ij} + 0.5}{Y_{ij+1} + \ldots + Y_{ik} + 0.5}$$
Testing for proportional odds

Assume that counts for the $i$th factor combination are from independent

$$(Y_{i1}, \ldots, Y_{i5}) \sim \text{Mult}(18, (\pi_{i1}, \ldots, \pi_{i5}))$$

**Proportional odds model**

$$\log \frac{\pi_{i1} + \ldots + \pi_{ij}}{\pi_{ij+1} + \ldots + \pi_{i5}} = \alpha_j - \beta w_i - \delta z_i$$

where $w_i$ is 0 (cold) or 1 (warm), $z_i$ is 0 (no) or 1 (yes),

$\beta, \delta \in \mathbb{R}$, $\alpha_1 < \ldots < \alpha_4 < \alpha_5 = \infty$

---

$^2$see, McCullagh (1980)  
$^3$see, Peterson and Harrell (1990)
Testing for proportional odds

Assume that counts for the $i$th factor combination are from independent

$$(Y_{i1}, \ldots, Y_{i5}) \sim \text{Mult}(18, (\pi_{i1}, \ldots, \pi_{i5}))$$

**Proportional odds model**

$$\log \frac{\pi_{i1} + \ldots + \pi_{ij}}{\pi_{ij+1} + \ldots + \pi_{i5}} = \alpha_j - \beta w_i - \delta z_i$$

where $w_i$ is 0 (cold) or 1 (warm), $z_i$ is 0 (no) or 1 (yes), $\beta, \delta \in \mathbb{R}$, $\alpha_1 < \ldots < \alpha_4 < \alpha_5 = \infty$

**Partial proportional odds model**

$$\log \frac{\pi_{i1} + \ldots + \pi_{ij}}{\pi_{ij+1} + \ldots + \pi_{i5}} = \alpha_j - \gamma_j w_i - \delta z_i$$

---

$^2$see, McCullagh (1980)

$^3$see, Peterson and Harrell (1990)
Testing for proportional odds

Assume that counts for the $i$th factor combination are from independent

$$(Y_{i1}, \ldots, Y_{i5}) \sim \text{Mult}(18, (\pi_{i1}, \ldots, \pi_{i5}))$$

**Proportional odds model**\(^2\)

$$\log \frac{\pi_{i1} + \ldots + \pi_{ij}}{\pi_{ij+1} + \ldots + \pi_{i5}} = \alpha_j - \beta w_i - \delta z_i$$

where $w_i$ is 0 (cold) or 1 (warm), $z_i$ is 0 (no) or 1 (yes), $\beta, \delta \in \mathbb{R}$, $\alpha_1 < \ldots < \alpha_4 < \alpha_5 = \infty$

**Partial proportional odds model**\(^3\)

$$\log \frac{\pi_{i1} + \ldots + \pi_{ij}}{\pi_{ij+1} + \ldots + \pi_{i5}} = \alpha_j - \gamma_j w_i - \delta z_i$$

Proportional odds hypothesis $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \beta$

---

\(^2\)see, McCullagh (1980)

\(^3\)see, Peterson and Harrell (1990)
Testing for proportional odds

Use Wald statistic

\[(L\hat{\gamma})^T \left\{ LF\gamma\gamma(\hat{\theta})L^T \right\}^{-1} L\hat{\gamma}\]

with a \(\chi^2_3\) limiting distribution under proportional odds

\(F\gamma\gamma(\theta)\) is \(\gamma\)-block of the inverse Fisher information matrix

\(L\) is a matrix of \(\gamma\)-contrasts

\[
\begin{bmatrix}
1 & . & . & -1 \\
. & 1 & . & -1 \\
. & . & 1 & -1
\end{bmatrix}
\]

\(^5\)see, Pratt (1981) and Agresti (2010, §3.4.5) for sufficient conditions
Testing for proportional odds

Use Wald statistic

\[(L\hat{\gamma})^T \left\{ LF\gamma\gamma(\hat{\theta})L^T \right\}^{-1} L\hat{\gamma}\]

with a \(\chi^2_3\) limiting distribution under proportional odds

\(F\gamma\gamma(\theta)\) is \(\gamma\)-block of the inverse Fisher information matrix

\(L\) is a matrix of \(\gamma\)-contrasts

\[
\begin{pmatrix}
1 & . & -1 \\
. & 1 & -1 \\
. & . & 1 \\
\end{pmatrix}
\]

Maximum likelihood\(^4\) returns infinite estimates\(^5\)

<table>
<thead>
<tr>
<th>(\alpha_1)</th>
<th>(\alpha_2)</th>
<th>(\alpha_3)</th>
<th>(\alpha_4)</th>
<th>(\gamma_1)</th>
<th>(\gamma_2)</th>
<th>(\gamma_3)</th>
<th>(\gamma_4)</th>
<th>(\delta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.27</td>
<td>1.10</td>
<td>3.77</td>
<td>24.90</td>
<td>21.10</td>
<td>2.15</td>
<td>2.87</td>
<td>22.55</td>
<td>1.47</td>
</tr>
<tr>
<td>Maximum absolute log-likelihood gradient: (10^{-6})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| -1.27 | 1.10 | 3.77 | 33.89 | 30.10 | 2.15 | 2.87 | 31.55 | 1.47 |
| Maximum absolute log-likelihood gradient: \(10^{-10}\) |

\(^4\)estimation here is done using the R package ordinal (Christensen, 2015)

\(^5\)see, Pratt (1981) and Agresti (2010, §3.4.5) for sufficient conditions
Requirements from a good estimator for PO models

Same or similar properties with the MLE (e.g. asymptotic efficiency)

Finite estimates and corresponding standard errors

Invariance to data (dis)aggregation

<table>
<thead>
<tr>
<th>Contact</th>
<th>Temp</th>
<th>Rating</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1  2  3  4  5</td>
<td>1  2  3  4  5</td>
</tr>
<tr>
<td>No</td>
<td>Cold</td>
<td>4  9  5  0  0</td>
<td>4  9  5  0  0</td>
</tr>
<tr>
<td></td>
<td>Warm</td>
<td>0  5  8  3  2</td>
<td>0  4  6  1  2</td>
</tr>
<tr>
<td>Yes</td>
<td>Cold</td>
<td>1  7  8  2  0</td>
<td>1  7  8  2  0</td>
</tr>
<tr>
<td></td>
<td>Warm</td>
<td>0  1  5  7  5</td>
<td>0  1  5  7  5</td>
</tr>
</tbody>
</table>

Optimal sampling properties which are preserved under linear parameter transformations (e.g. \( L \) contrasts, reversal of categories and so on)
Outline

1. Testing for proportional odds
2. Reducing bias
3. Direction of shrinkage
4. Discussion
Cumulative link model$^6$

Vectors of counts on $k$ ordered categories are from independent multinomial vectors $Y_1, \ldots, Y_n$ with

$$Y_i \mid x_i \sim \text{Mult}(m_i, (\pi_{i1}, \ldots, \pi_{ik}))$$

$$g(\pi_{i1} + \ldots + \pi_{ij}) = \alpha_j + \beta^T x_i = \sum_{t=1}^{p+k-1} \theta_t z_{ijt}$$

$x_i$ is a $p$-vector of explanatory variables

$\alpha_1 < \ldots < \alpha_{k-1} < \alpha_k = \infty$ and $\beta \in \mathbb{R}^p$

$\theta = (\alpha_1, \ldots, \alpha_{k-1}, \beta_1, \ldots, \beta_p)^T$

$g(.)$ is a monotone increasing, differentiable link function

Special cases

Proportional odds model: $g = \text{logit}$

Proportional hazards model (grouped survival times): $g = \text{cloglog}$

$^6$see, McCullagh (1980) and Agresti (2010, §5.1)
Bias reduction through adjusted score functions

Maximum likelihood estimator

\[ \hat{\theta} \leftarrow \left\{ \sum_{i} \sum_{j=1}^{k-1} g'_{ij} \left( \frac{Y_{ij} - Y_{ij+1}}{\pi_{ij}} \right) z_{ijt} = 0 \right\} \]

where \( g'_{ij} = dg^{-1}(\eta)/d\eta \)

Bias-reduced estimator\(^7\)

An estimator with smaller asymptotic bias than \( \hat{\theta} \) is

\[ \theta^* \leftarrow \left\{ \sum_{i} \sum_{j=1}^{k-1} g'_{ij} \left( \frac{y_{ij}^* + c_{ij} - c_{ij-1}}{\pi_{ij}} - \frac{Y_{ij+1} + c_{ij+1} - c_{ij}}{\pi_{ij+1}} \right) z_{ijt} = 0 \right\} \]

where \( c_{ij} = m_i g''_{ij} [Z_i F^{-1} Z_i^T]_{jj}/2 \) and \( c_{i0} = c_{ik} = 0 \)

\(^7\)see, K. (2014, RSSB) and K. and Firth (2009, B'ka) for method
Iterative maximum likelihood fits

The kernel in the adjusted score (omitting $i$) is

$$\frac{y_j + d_j}{\pi_j} - \frac{y_{j+1} + d_{j+1}}{\pi_{j+1}}$$

where $d_j = c_j - c_{j-1}$
Iterative maximum likelihood fits

The kernel in the adjusted score (omitting $i$) is

$$\frac{y_j + d_j}{\pi_j} - \frac{y_{j+1} + d_{j+1}}{\pi_{j+1}}$$

where $d_j = c_j - c_{j-1}$

**Empirical cumulative logits**

$$\log \frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_k} = \alpha_j$$

$d_1 = 0.5 - \pi_1$, $d_j = -\pi_j$ ($j = 2, \ldots, k - 1$), and $d_k = 0.5 - \pi_k$

1. add 0.5 to the counts of the first and last category only
2. use ML on the adjusted data

The bias-reduced estimators end up being the empirical cumulative logits

$$\alpha_j^* = \log \frac{Y_1 + \ldots + Y_j + 0.5}{Y_{j+1} + \ldots + Y_k + 0.5}$$
Iterative maximum likelihood fits

The kernel in the adjusted score (omitting $i$) is

$$\frac{y_j + d_j}{\pi_j} - \frac{y_{j+1} + d_{j+1}}{\pi_{j+1}}$$

where $d_j = c_j - c_{j-1}$

More general models

The kernel can be re-expressed as

$$\frac{y_j + d_j I_j - \pi_j d_{j+1} (1 - l_{j+1})/\pi_{j+1}}{\pi_j} - \frac{y_{j+1} + d_{j+1} l_{j+1} - \pi_{j+1} d_j (1 - l_j)/\pi_j}{\pi_{j+1}}$$

where $l_j$ is 1 if $d_j > 0$ and 0 else

Iterative maximum likelihood fits

At the $u$th iteration

1. add $d_j^{(u)} l_j^{(u)} - \pi_j^{(u)} d_{j+1}^{(u)} (1 - l_{j+1}^{(u)})/\pi_{j+1}^{(u)}$ to $y_j$

2. fit the model on the adjusted counts with maximum likelihood
Properties of bias-reduced estimator

\( \theta^* \) is equivariant under linear transformations\(^8\)

i.e. the bias-reduced estimator of \( L\theta \) is \( L\theta^* \)

\(^8\text{see, K. (2014, RSSB, §6-7) for proofs}\)
Properties of bias-reduced estimator

\( \theta^* \) is equivariant under linear transformations\(^8\)

\( \theta^* \) and \( \hat{\theta} \) have the same asymptotic distribution, i.e. \( N(\theta, F^{-1}(\theta)) \)\(^9\)

First-order inference tools, like Wald tests, apply unaltered

Standard errors and estimated variance-covariance matrices, in general, can be computed using \( F^{-1}(\theta^*) \)

---

\(^8\)see, K. (2014, RSSB, §6-7) for proofs  
\(^9\)see, Firth (1993) and K. and Firth (2009)
Properties of bias-reduced estimator

$\theta^*$ is equivariant under linear transformations$^8$

$\theta^*$ and $\hat{\theta}$ have the same asymptotic distribution, i.e. $N(\theta, F^{-1}(\theta))$$^9$

$\theta^*$ has always finite components

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
<th>$\gamma_1$</th>
<th>$\gamma_2$</th>
<th>$\gamma_3$</th>
<th>$\gamma_4$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimates</strong></td>
<td>-1.27</td>
<td>1.10</td>
<td>3.77</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>2.15</td>
<td>2.87</td>
<td>$\infty$</td>
<td>1.47</td>
</tr>
<tr>
<td><strong>Max. likelihood</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Std. errors</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
<th>$\gamma_1$</th>
<th>$\gamma_2$</th>
<th>$\gamma_3$</th>
<th>$\gamma_4$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimates</strong></td>
<td>-1.19</td>
<td>1.05</td>
<td>3.50</td>
<td>5.20</td>
<td>2.62</td>
<td>2.05</td>
<td>2.65</td>
<td>2.96</td>
<td>1.40</td>
</tr>
<tr>
<td><strong>Bias reduction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Std. errors</strong></td>
<td>0.50</td>
<td>0.44</td>
<td>0.74</td>
<td>1.47</td>
<td>1.52</td>
<td>0.58</td>
<td>0.75</td>
<td>1.50</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Testing for proportional odds using $\hat{\theta}$

$W = 0.7502$ leading to a $p$-value of 0.861 (based on $\chi^2_3$)

$^8$see, K. (2014, RSSB, §6-7) for proofs

$^9$see, Firth (1993) and K. and Firth (2009)
Properties of bias-reduced estimator

\( \theta^* \) is equivariant under linear transformations\(^8\)

\( \theta^* \) and \( \hat{\theta} \) have the same asymptotic distribution, i.e. \( N(\theta, F^{-1}(\theta)) \)\(^9\)

\( \theta^* \) has always finite components

\( \theta^* \) is invariant to data (dis)aggregation

<table>
<thead>
<tr>
<th>contact</th>
<th>temp</th>
<th>rating</th>
<th>rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>no</td>
<td>cold</td>
<td>4 9 5 0 0</td>
<td>4 9 5 0 0</td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>0 5 8 3 2</td>
<td>0 4 6 1 2</td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>0 1 2 2 0</td>
<td>0 1 2 2 0</td>
</tr>
<tr>
<td>yes</td>
<td>cold</td>
<td>1 7 8 2 0</td>
<td>1 7 8 2 0</td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>0 1 5 7 5</td>
<td>0 1 5 7 5</td>
</tr>
</tbody>
</table>

Adding constants + ML is dangerous for general models

\(^8\)see, K. (2014, RSSB, §6-7) for proofs

\(^9\)see, Firth (1993) and K. and Firth (2009)
Graduate admissions in Stanford U

Data

Admission scores and candidate characteristics from 106 applications to the political science PhD at Stanford University

rater’s score ($1 < 2 < 3 < 4 < 5$)
interest in American politics and political theory ($z_{i1}$ and $z_{i2}$; $1$:yes, $0$:no)
standardized score on quantitative and verbal parts of GRE ($x_{i1}$ and $x_{i2}$)
gender ($g_i$; $0$:male and $1$:female)

Proportional odds model

$$\text{logit}(\pi_{i1} + \ldots + \pi_{ij}) = \alpha_j - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 z_{i1} - \beta_4 z_{i2} - \beta_5 g_i$$

ML estimates

$\hat{\beta}_1 = 1.993$, $\hat{\beta}_2 = 0.892$, $\hat{\beta}_3 = 2.816$, $\hat{\beta}_4 = 0.009$, $\hat{\beta}_5 = 1.215$

---

Supplementary material: rater F1 in the analysis in Jackman (2004); R package pscl (Jackman, 2015)
## Simulation results

<table>
<thead>
<tr>
<th></th>
<th>Bias</th>
<th>MSE</th>
<th>$\text{Bias}^2 / \text{Variance}$ (%)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ML</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.13</td>
<td>0.14</td>
<td>13.90</td>
<td>94.42</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.05</td>
<td>0.06</td>
<td>5.02</td>
<td>94.15</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.22</td>
<td>0.79</td>
<td>6.29</td>
<td>94.68</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.00</td>
<td>0.64</td>
<td>0.00</td>
<td>94.50</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0.07</td>
<td>0.24</td>
<td>2.33</td>
<td>94.21</td>
</tr>
<tr>
<td><strong>BR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.00</td>
<td>0.11</td>
<td>0.00</td>
<td>95.05</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>95.09</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.01</td>
<td>0.59</td>
<td>0.01</td>
<td>95.32</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.00</td>
<td>0.56</td>
<td>0.00</td>
<td>95.55</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>94.99</td>
</tr>
</tbody>
</table>

*figures are based on 10000 samples under the maximum likelihood fit*
Outline

1  Testing for proportional odds

2  Reducing bias

3  Direction of shrinkage

4  Discussion
Direction of shrinkage

Model is "shrunken" to a binomial GLM for the boundary categories

**Demonstration**

Complete enumeration (3136) of tables of the form

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

Model: \(g(\pi_{i1} + \ldots + \pi_{ij}) = \alpha_j - \beta x_i\)

Calculate fitted probabilities based on \(\hat{\theta}\) and \(\theta^*\) for each table and for \(g = \text{logit}\) and \(g = \text{cloglog}\).
BR probabilities for intermediate categories tend to shrink to 0
BR probabilities for 1st (6th) category tend to shrink to $g^{-1}(0)$ (1 - $g^{-1}(0)$)
Outline

1  Testing for proportional odds
2  Reducing bias
3  Direction of shrinkage
4  Discussion
Discussion I

**Estimation properties**
\( \theta^* \) has all the required properties when estimating cumulative link models and is always finite.

First-order likelihood inference applies in a “plug-in” fashion.

**Shrinkage**

Model is shrunken towards a binomial GLM for the boundary categories.

Adjusted scores provide just enough regularization to correct for bias and improve inference. Different regularization schemes may be needed for other tasks (e.g. prediction).

**Confidence intervals**

When testing for extreme effects, default tests (e.g. Wald or adjusted score) always reject due to the interplay of finiteness and discreteness.
Discussion II

Software


handles general models and will soon be part of the brglm2 R package

References


