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

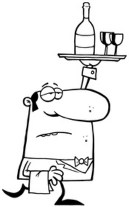
31 August 2017
CEN ISBS 2017 Joint Conference
Vienna, Austtria

Outline

Outline

Wine tasting data[1]

		rating				
contact	temp	1	2	3	4	5
no	cold	4	9	5	0	0
	warm	0	5	8	3	2
yes	cold	1	7	8	2	0
	warm	0	1	5	7	5

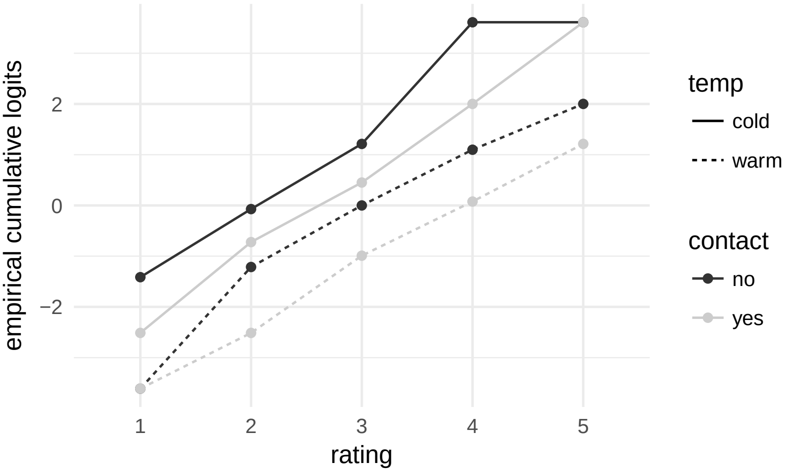


Experiment on the effect of factors on the bitterness of white wine

contact of juice with skin and temperature 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 ith 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),
$\beta, \delta \in \Re$, $\alpha_1 < \ldots < \alpha_4 < \alpha_5 = \infty$

[2]see, McCullagh (1980)
[3]see, Peterson and Harrell (1990)

Testing for proportional odds

Assume that counts for the ith 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), $\beta, \delta \in \Re$, $\alpha_1 < \ldots < \alpha_4 < \alpha_5 = \infty$

Partial proportional odds model[3]

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

---

[2] see, McCullagh (1980)
[3] see, Peterson and Harrell (1990)

Testing for proportional odds

Assume that counts for the ith 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),
$\beta, \delta \in \Re$, $\alpha_1 < \ldots < \alpha_4 < \alpha_5 = \infty$

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

$$(L\hat{\gamma})^\top \left\{ L F^{\gamma\gamma}(\hat{\theta}) L^\top \right\}^{-1} L\hat{\gamma}$$

with a χ^2_3 limiting distribution under proportional odds

$F^{\gamma\gamma}(\theta)$ is γ-block of the inverse Fisher information matrix

L is a matrix of γ-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

$$(L\hat{\gamma})^\top \left\{ L F^{\gamma\gamma}(\hat{\theta}) L^\top \right\}^{-1} L\hat{\gamma}$$

with a χ^2_3 limiting distribution under proportional odds

$F^{\gamma\gamma}(\theta)$ is γ-block of the inverse Fisher information matrix

L is a matrix of γ-contrasts $\begin{bmatrix} 1 & . & . & -1 \\ . & 1 & . & -1 \\ . & . & 1 & -1 \end{bmatrix}$

Maximum likelihood[4] returns infinite estimates[5]

α_1	α_2	α_3	α_4	γ_1	γ_2	γ_3	γ_4	δ
-1.27	1.10	3.77	24.90	21.10	2.15	2.87	22.55	1.47
Maximum absolute log-likelihood gradient: 10^{-6}								
-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

		Aggregated					Disaggregated				
		rating									
contact	temp	1	2	3	4	5	1	2	3	4	5
no	cold	4	9	5	0	0	4	9	5	0	0
	warm	0	5	8	3	2	0	4	6	1	2
	warm						0	1	2	2	0
yes	cold	1	7	8	2	0	1	7	8	2	0
	warm	0	1	5	7	5	0	1	5	7	5

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

Outline

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 $\beta \in \Re^p$

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

$g(.)$ 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

$$\hat{\theta} \leftarrow \left\{ \sum_i \sum_{j=1}^{k-1} g'_{ij} \left(\frac{y_{ij}}{\pi_{ij}} - \frac{y_{ij+1}}{\pi_{ij+1}} \right) z_{ijt} = 0 \right\}$$

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

Bias-reduced estimator[7]

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

$$\theta^* \leftarrow \left\{ \sum_i \sum_{j=1}^{k-1} g'_{ij} \left(\frac{\overbrace{y_{ij} + c_{ij} - c_{ij-1}}^{\text{adjusted response } y^*_{ij}}}{\pi_{ij}} - \frac{y_{ij+1} + c_{ij+1} - c_{ij}}{\pi_{ij+1}} \right) z_{ijt} = 0 \right\}$$

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

$$\frac{y_j + \overbrace{d_j I_j - \pi_j d_{j+1}(1 - I_{j+1})/\pi_{j+1}}^{\text{always} \geq 0}}{\pi_j} - \frac{y_{j+1} + d_{j+1} I_{j+1} - \pi_{j+1} d_j (1 - I_j)/\pi_j}{\pi_{j+1}}$$

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

Iterative maximum likelihood fits

At the uth iteration

1. add $d_j^{(u)} I_j^{(u)} - \pi_j^{(u)} d_{j+1}^{(u)}(1 - I_{j+1}^{(u)})/\pi_{j+1}^{(u)}$ to y_j
2. fit the model on the adjusted counts with maximum likelihood

Properties of bias-reduced estimator

θ^* is equivariant under linear transformations[8]

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

[8] see, K. (2014, RSSB, §6-7) for proofs

Properties of bias-reduced estimator

θ^* is equivariant under linear transformations[8]

θ^* 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

θ^* is equivariant under linear transformations[8]

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

θ^* has always finite components

	α_1	α_2	α_3	α_4	γ_1	γ_2	γ_3	γ_4	δ
					Maximum likelihood				
Estimates	-1.27	1.10	3.77	∞	∞	2.15	2.87	∞	1.47
Std. errors	-	-	-	-	-	-	-	-	-
					Bias reduction				
Estimates	-1.19	1.05	3.50	5.20	2.62	2.05	2.65	2.96	1.40
Std. errors	0.50	0.44	0.74	1.47	1.52	0.58	0.75	1.50	0.46

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

W $= 0.7502$ leading to a p-value of 0.861 (based on χ^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

θ^* is equivariant under linear transformations[8]

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

θ^* has always finite components

θ^* is invariant to data (dis)aggregation

		Aggregated					Disaggregated				
		rating									
contact	temp	1	2	3	4	5	1	2	3	4	5
no	cold	4	9	5	0	0	4	9	5	0	0
	warm	0	5	8	3	2	0	4	6	1	2
	warm						0	1	2	2	0
yes	cold	1	7	8	2	0	1	7	8	2	0
	warm	0	1	5	7	5	0	1	5	7	5

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

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

[10] rater F1 in the analysis in Jackman (2004); R package pscl (Jackman, 2015)

Simulation results

		Bias	MSE	Bias^2/Variance (%)	Coverage (%)
ML	β_1	0.13	0.14	13.90	94.42
	β_2	0.05	0.06	5.02	94.15
	β_3	0.22	0.79	6.29	94.68
	β_4	0.00	0.64	0.00	94.50
	β_5	0.07	0.24	2.33	94.21
BR	β_1	0.00	0.11	0.00	95.05
	β_2	0.00	0.05	0.00	95.09
	β_3	0.01	0.59	0.01	95.32
	β_4	0.00	0.56	0.00	95.55
	β_5	-0.00	0.21	0.00	94.99

figures are based on 10000 samples under the maximum likelihood fit

Outline

Direction of shrinkage

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

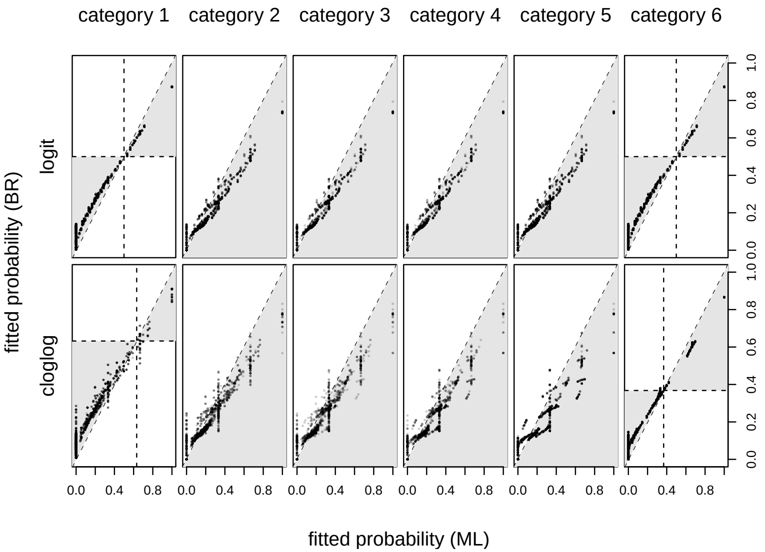
Demonstration

Complete enumeration (3136) of tables of the form

	category						
x	1	2	3	4	5	6	total
-0.5							3
0.5							3

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

Calculate fitted probabilities based on $\hat{\theta}$ and θ^* for each table and for $g = \mathrm{logit}$ and $g = \mathrm{cloglog}$.



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

Discussion I

Estimation properties

θ^* 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

bpolr R function in the supplementary material of

Kosmidis (2014). Improved estimation in cumulative link models.
Journal of the Royal Statistical Society: Series B, 76
[DOI: 10.1111/rssb.12025]

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

Kosmidis (2017). brglm2: Bias reduction in generalized linear models.
R package version 0.1.4
[URL: https://cran.r-project.org/package=brglm2]

References I

Agresti (2010). *Analysis of Ordinal Categorical Data (2nd Edition)*. John Wiley & Sons.

Christensen, R. H. B. (2015). ordinal—regression models for ordinal data. R package version 2015.6-28. https://cran.r-project.org/package=ordinal.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *80*(1), 27–38.

Jackman, S. (2004). What do we learn from graduate admissions committees? a multiple rater, latent variable model, with incomplete discrete and continuous indicators. *Political Analysis 12*(4), 400–424.

Jackman, S. (2015). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Stanford, California: Department of Political Science, Stanford University. R package version 1.4.9.

K. (2014). Improved estimation in cumulative link models. *Journal of the Royal Statistical Society, Series B 76*(1), 169–196.

K. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika 96*(4), 793–804.

McCullagh, P. (1980). Regression models for ordinal data. *42*, 109–142.

Peterson, B. and J. Harrell, Frank E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics 39*, 205–217.

Pratt, J. W. (1981). Concavity of the log likelihood (Corr: V77 p954). *Journal of the American Statistical Association 76*, 103–106.

Randall (1989). The analysis of sensory data by generalised linear model. *Biometrical Journal 7*, 781–793.