Regularized Ordinal Regression and the ordinalNet R Package

Bret Hanlon, University of Wisconsin-Madison

April 6, 2022

Acknowledgements-1

- Mike Wurm, Data Science, Google
- Paul Rathouz, Director of the Biomedical Science Data Hub, University of Texas-Austin
- dissertation. Wurm, Michael J. Regression Models and Optimization Techniques for Ordinal, Survival, and Exponential Family Distributions. The University of Wisconsin-Madison, 2017.
- manuscript. Wurm, M. J., Rathouz, P. J., & Hanlon, B. M. (2021). Regularized Ordinal Regression and the ordinalNet R Package. Journal of Statistical Software, 99(6), 1–42.

Acknowledgements-2

- Wilhelm et al's elucidation of iteratively reweighted least squares (IRLS) for multinomial models.

  Wilhelm MS, Carter EM, Hubert JJ. Multivariate iteratively re-weighted least squares, with applications to dose-response data. Environmetrics, 9(3):303–315.

- Yee's vector generalized additive models (VGAM).

  Yee TW. Vector generalized linear and additive models: with an implementation in R. New York: Springer 2015.

  Yee TW. The VGAM package for categorical data analysis. Journal of Statistical Software. 2010 Jan 5;32:1-34.

Motivating example

Motivating data from The Cancer Genome Atlas

- Archer et al. (2014, 2010)
- 56 subjects
  - 20 normal liver tissue (healthy)
  - 16 cirrhotic tissue (disease)
  - 20 with hepatocellular carcinoma (severe disease)
- Methylation levels for 45 genes
- Goal: build predictive model to classify tissue based on methylation profile
- $p \approx n \Rightarrow$ predictive model requires regularization
- Excerpt from data:

Liver	CDKN2B_seq_50_S294_F	DDIT3_P1313_R	ERN1_P809_R	GML_E144_F	. . .
Normal	-0.35633416	-1.31056556	0.9657593169	0.45453048	
Normal	-0.47508349	-0.66903874	0.9537646464	0.57388473	
Tumor	1.37096496	1.37428021	0.4118315292	0.72121947	
Tumor	0.54953731	1.06033861	0.4063716945	-1.08468387	
Cirrhosis non-HCC	0.55573061	0.59576393	-0.6224016200	0.70409687	
Cirrhosis non-HCC	-1.17830718	0.03673283	0.0004964171	-1.14753553	
⋮					

Ordinal regression overview

- Response variable has finite number of ordered categories
    - For example, 1=poor, 2=fair, 3=good, 4=excellent
- Common approaches
    - Treat outcome as numeric
        - Interpretation is problematic
    - Multinomial regression
        - Does not take advantage of label ordering
    - Ordinal regression
        - Fewer parameters than multinomial regression, but less flexible.

Motivation & research contributions

- **Motivation 1:** Limited software available for ordinal regression regularization and variable selection (true in 2015).
  - **Contribution:** Proposed and implemented coordinate descent algorithm for broad class of models with elastic net penalty

Motivation & research contributions

- **Motivation 1:** Limited software available for ordinal regression regularization and variable selection (true in 2015).
  - **Contribution:** Proposed and implemented coordinate descent algorithm for broad class of models with elastic net penalty
- **Motivation 2:** How to choose between ordinal and unordered multinomial models for ordinal data?
  - **Contribution:** Developed model parameterization that blends ordinal and multinomial regression.

Examples of common multinomial & ordinal models

- Multinomial logistic regression
  - $P(Y = m | X = x) = \frac{\exp(\alpha_m + x^\top \beta_m)}{1 + \sum_{k=1}^{K-1} \exp(\alpha_k + x^\top \beta_k)}$
  - $K - 1$ sets of coefficients (β_k) and intercepts (α_k).
  - Invariant to class label ordering.

Examples of common multinomial & ordinal models

- Multinomial logistic regression
  - $P(Y = m|X = x) = \frac{\exp(\alpha_m + x^\top \beta_m)}{1 + \sum_{k=1}^{K-1} \exp(\alpha_k + x^\top \beta_k)}$
  - $K - 1$ sets of coefficients (β_k) and intercepts (α_k).
  - Invariant to class label ordering.

- Proportional odds model
  - $P(Y \leq m|X = x) = \text{logit}^{-1}(\alpha_m + x^\top \beta)$
  - Single set of coefficients (β) and $K - 1$ intercepts (α_k).
  - The linear combination $x^\top \beta$ shifts all cumulative probabilities up or down. **This is the defining characteristic of an ordinal model!**

Examples of common multinomial & ordinal models

- Multinomial logistic regression
  - $P(Y = m|X = x) = \frac{\exp(\alpha_m + x^\top \beta_m)}{1 + \sum_{k=1}^{K-1} \exp(\alpha_k + x^\top \beta_k)}$
  - $K - 1$ sets of coefficients (β_k) and intercepts (α_k).
  - Invariant to class label ordering.

- Proportional odds model
  - $P(Y \leq m|X = x) = \text{logit}^{-1}(\alpha_m + x^\top \beta)$
  - Single set of coefficients (β) and $K - 1$ intercepts (α_k).
  - The linear combination $x^\top \beta$ shifts all cumulative probabilities up or down. **This is the defining characteristic of an ordinal model!**

- Goal: design a blended model with the large model space of multinomial logistic that can be penalized toward an ordinal model.

Proportional odds model: latent variable interpretation

Suppose the outcome has 3 categories. Then for $m \in \{1, 2\}$,

$$P(Y \leq m | X = x) = \text{logit}^{-1}(\alpha_m + x^\top \beta) = F_Z(\alpha_m + x^\top \beta) \ ,$$

where F_Z is the cdf of a standard logistic distribution.

Proportional odds model: latent variable interpretation

Suppose the outcome has 3 categories. Then for $m \in \{1, 2\}$,

$$P(Y \leq m | X = x) = \text{logit}^{-1}(\alpha_m + x^\top \beta) = F_Z(\alpha_m + x^\top \beta) \ ,$$

where F_Z is the cdf of a standard logistic distribution.

$Y = 1$	$Y = 2$	$Y = 3$
$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$
$\boldsymbol{Z} < \alpha_1 + x^\top \beta$	$\alpha_1 + x^\top \beta < \boldsymbol{Z} < \alpha_2 + x^\top \beta$	$\alpha_2 + x^\top \beta < \boldsymbol{Z}$

Proportional odds model: latent variable interpretation

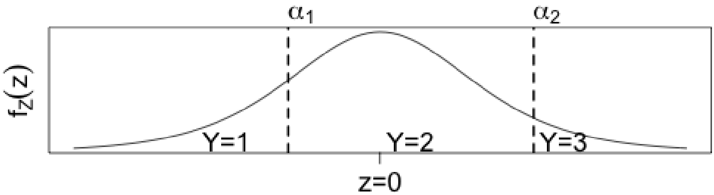
Suppose the outcome has 3 categories. Then for $m \in \{1, 2\}$,

$$P(Y \leq m | X = x) = \text{logit}^{-1}(\alpha_m + x^\top \beta) = F_Z(\alpha_m + x^\top \beta) ,$$

where F_Z is the cdf of a standard logistic distribution.

$Y = 1$	$Y = 2$	$Y = 3$
$\Longleftrightarrow$	$\Longleftrightarrow$	$\Longleftrightarrow$
$\boldsymbol{Z} < \alpha_1 + x^\top \beta$	$\alpha_1 + x^\top \beta < \boldsymbol{Z} < \alpha_2 + x^\top \beta$	$\alpha_2 + x^\top \beta < \boldsymbol{Z}$

At $x = 0$, the intercepts determine category cut points:



Proportional odds model: latent variable interpretation

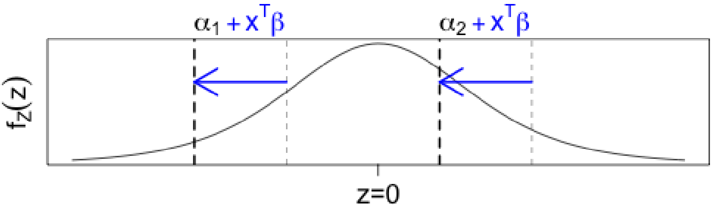
Suppose the outcome has 3 categories. Then for $m \in \{1, 2\}$,

$$P(Y \leq m | X = x) = \text{logit}^{-1}(\alpha_m + x^\top \beta) = F_Z(\alpha_m + x^\top \beta) \ ,$$

where F_Z is the cdf of a standard logistic distribution.

$Y = 1$ $\Longleftrightarrow$ $\boldsymbol{Z} < \alpha_1 + x^\top \beta$	$Y = 2$ $\Longleftrightarrow$ $\alpha_1 + x^\top \beta < \boldsymbol{Z} < \alpha_2 + x^\top \beta$	$Y = 3$ $\Longleftrightarrow$ $\alpha_2 + x^\top \beta < \boldsymbol{Z}$

$\boldsymbol{x^\top \beta}$ **shifts probability toward higher or lower categories:**

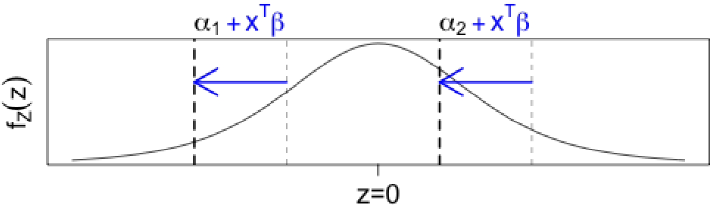


Generalizing the Proportional Odds Model

$$\text{logit}\Big(P(Y \le m|X = x)\Big) = \dots$$

- Proportional odds

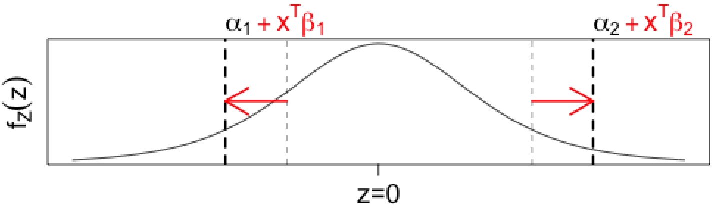
$$= \alpha_m + x^\top \beta$$



Generalizing the Proportional Odds Model

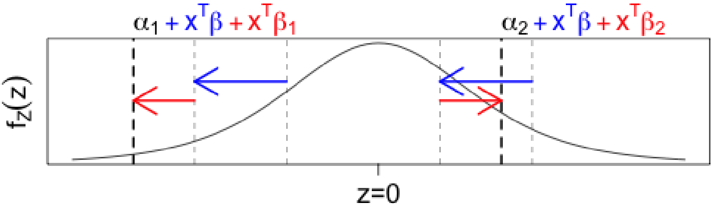
$$\text{logit}\Big(P(Y \leq m | X = x)\Big) = \dots$$

- Proportional odds
  $= \alpha_m + x^\top \beta$
- Partial proportional odds (Peterson et al., 1990)
  $= \alpha_m + x^\top \beta_m$



$$\text{logit}\Big(P(Y \leq m | X = x)\Big) = \dots$$

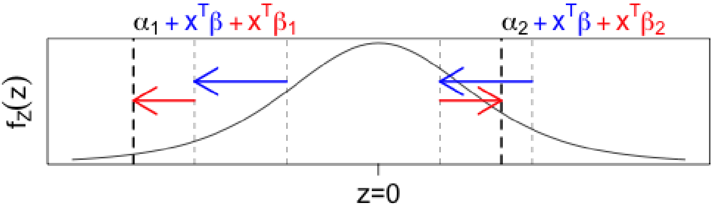
- Proportional odds
  $= \alpha_m + x^\top \beta$
- Partial proportional odds (Peterson et al., 1990)
  $= \alpha_m + x^\top \beta_m$
- Blended model (Wurm et al., 2017b)
  $= \alpha_m + x^\top \beta + x^\top \beta_m$



Generalizing the Proportional Odds Model

$$\text{logit}\Big(P(Y \le m | X = x)\Big) = \dots$$

- Proportional odds
  $= \alpha_m + x^\top \beta \leftarrow$ **"Parallel"**
- Partial proportional odds (Peterson et al., 1990)
  $= \alpha_m + x^\top \beta_m \leftarrow$ **"Nonparallel"**
- Blended model (Wurm et al., 2017b)
  $= \alpha_m + x^\top \beta + x^\top \beta_m \leftarrow$ **"Semi-parallel"**



Generalizing the Proportional Odds Model

$$\text{logit}\Big(P(Y \leq m|X = x)\Big) = \dots$$

- Proportional odds

  $= \alpha_m + x^\top \beta \leftarrow$ **"Parallel"**

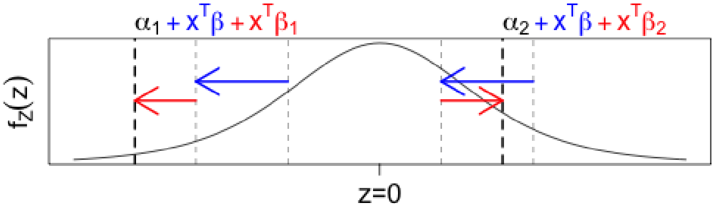
- Partial proportional odds (Peterson et al., 1990)

  $= \alpha_m + x^\top \beta_m \leftarrow$ **"Nonparallel"**

- Blended model (Wurm et al., 2017b)

  $= \alpha_m + x^\top \beta + x^\top \beta_m \leftarrow$ **"Semi-parallel"**

  - Elastic net penalty can be applied to both β and β_m, setting unimportant coefficients to zero.



Generalizing the Proportional Odds Model

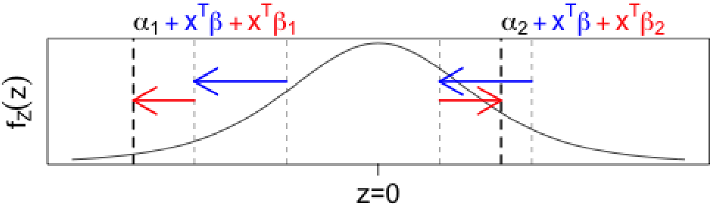
$$\text{logit}\Big(P(Y \leq m|X = x)\Big) = \dots$$

- Proportional odds

  $= \alpha_m + x^\top \beta \leftarrow$ **"Parallel"**
- Partial proportional odds (Peterson et al., 1990)

  $= \alpha_m + x^\top \beta_m \leftarrow$ **"Nonparallel"**
- Blended model (Wurm et al., 2017b)

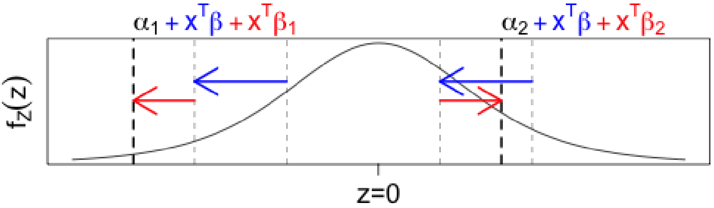
  $= \alpha_m + x^\top \beta + x^\top \beta_m \leftarrow$ **"Semi-parallel"**
  - Elastic net penalty can be applied to both β and β_m, setting unimportant coefficients to zero.
  - If parallel (ordinal) model is a good fit, then β terms will be kept.



Generalizing the Proportional Odds Model

$$\text{logit}\Big(P(Y \leq m | X = x)\Big) = \ldots$$

- Proportional odds
  $= \alpha_m + x^\top \beta \leftarrow$ **"Parallel"**
- Partial proportional odds (Peterson et al., 1990)
  $= \alpha_m + x^\top \beta_m \leftarrow$ **"Nonparallel"**
- Blended model (Wurm et al., 2017b)
  $= \alpha_m + x^\top \beta + x^\top \beta_m \leftarrow$ **"Semi-parallel"**
  - Elastic net penalty can be applied to both β and β_m, setting unimportant coefficients to zero.
  - If parallel (ordinal) model is a good fit, then β terms will be kept.
  - β_m terms will be kept as necessary to compensate for lack of fit.



ELMO

- The proportional odds model belongs to a broader class of models that have parallel/nonparallel/semi-parallel parameterizations.
- We call this the elementwise link multinomial-ordinal (ELMO).
- ELMO is a subset of Vector GLMs used in the VGAM R package (Yee and Wild, 1996; Yee, 2010, 2015).

ELMO (continued)

- Each ELMO model is defined by a family and an elementwise link function. Together, they determine a multivariate link function from class probabilities to linear combinations:

$$
\begin{pmatrix} P(Y=1|X=x) \\ \vdots \\ P(Y=K-1|X=x) \end{pmatrix} \overset{family}{\longrightarrow} \underbrace{\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_{K-1} \end{pmatrix}}_{\in(0,1)^{K-1}} \begin{matrix} \overset{link}{\rightarrow} \\ \rightarrow \\ \rightarrow \end{matrix} \overbrace{\underbrace{\begin{pmatrix} \alpha_1 + x^\top\beta + x^\top\beta_1 \\ \vdots \\ \alpha_{K-1} + x^\top\beta + x^\top\beta_{K-1} \end{pmatrix}}_{\in\mathbb{R}^{K-1}}}^{\text{Can be parallel/nonparallel/semi-parallel}}
$$

ELMO (continued)

- Each ELMO model is defined by a family and an elementwise link function. Together, they determine a multivariate link function from class probabilities to linear combinations:

$$\begin{pmatrix} P(Y=1|X=x) \\ \vdots \\ P(Y=K-1|X=x) \end{pmatrix} \xrightarrow{\textit{family}} \underbrace{\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_{K-1} \end{pmatrix}}_{\in(0,1)^{K-1}} \begin{matrix} \overset{\textit{link}}{\rightarrow} \\ \rightarrow \\ \rightarrow \end{matrix} \overbrace{\underbrace{\begin{pmatrix} \alpha_1 + x^\top\beta + x^\top\beta_1 \\ \vdots \\ \alpha_{K-1} + x^\top\beta + x^\top\beta_{K-1} \end{pmatrix}}_{\in\mathbb{R}^{K-1}}}^{\text{Can be parallel/nonparallel/semi-parallel}}$$

- For example, the proportional odds model:
  - The family specifies $\delta_k = P(Y \le k|X=x)$.
  - The link is logit.

ELMO (continued)

Family	δ_k
Cumulative Probability	$P(Y \leq k \vert X = x)$
Stopping Ratio	$P(Y = k \mid Y \geq k, X = x)$
Continuation Ratio	$P(Y > k \mid Y \geq k, X = x)$
Adjacent Category	$P(Y = k + 1 \mid k \leq Y \leq k + 1, X = x)$

The link function can be any binary regression link (e.g. logit, probit, complementary log-log).

ELMO (continued)

Family	δ_k
Cumulative Probability	$P(Y \leq k \vert X = x)$
Stopping Ratio	$P(Y = k \mid Y \geq k, X = x)$
Continuation Ratio	$P(Y > k \mid Y \geq k, X = x)$
Adjacent Category	$P(Y = k + 1 \mid k \leq Y \leq k + 1, X = x)$

The link function can be any binary regression link (e.g. logit, probit, complementary log-log).

Fun fact: The *unpenalized* adjacent category logit model is equivalent to multinomial logistic regression. Only the adjacent category parameterization has parallel/nonparallel/semi-parallel forms.

Elastic net penalty for semi-parallel model

- Penalized objective function is $-\frac{1}{n} \times \text{loglik} + \text{penalty}$
- Lasso penalty $= \lambda \left(\rho \|\beta\|_1 + \sum_{k=1}^{K-1} \|\beta_k\|_1 \right)$
  - $\lambda \geq 0$ determines overall penalty strength.
  - $\rho \geq 0$ determines penalty strength on ordinal coefficients.
    - Can be tuned, but $\rho = 1$ works well in practice.
- Ridge penalty $= \frac{\lambda}{2} \left(\rho \|\beta\|_2^2 + \sum_{k=1}^{K-1} \|\beta_k\|_2^2 \right)$
- Elastic net penalty $= \alpha \times$ Lasso penalty $+ (1 - \alpha) \times$ Ridge penalty
  - $\alpha \in [0, 1]$ determines weighting between lasso (L1) and ridge (L2) penalty.

Optimization algorithm

- Cyclic coordinate descent (Friedman et al., 2010, 2007).
- Algorithm applies whenever Fisher scoring algorithm for unpenalized model can be formulated as iteratively reweighted least squares (IRLS).
- ELMO models fit this framework (Wilhelm et al., 1998).
- Procedure:
  - Replace log-likelihood by its quadratic approximation of the form $-\frac{1}{2}\sum_{i=1}^{N}\|W_i^{1/2}(z_i - X_i\beta)\|^2$
  - Optimize approximated objective function marginally, one coefficient at a time. Cycle over coefficients until convergence.
  - Update quadratic approximation at new $\hat{\beta}$ estimate.

R Software

Software for lasso/elastic net penalty (no ordinal models)

- glmnet (Friedman et al., 2010)
- penalized (Goeman et al., 2017)

Software for ordinal logistic regression (no lasso/elastic net penalty)

- MASS::polr (Venables and Ripley, 2002)
- rms::lrm (Harrell, 2015)
- ordinalgmifs (Archer et al., 2014)
  GMIFS = Generalized Monotone Incremental Forward Stagewise regression

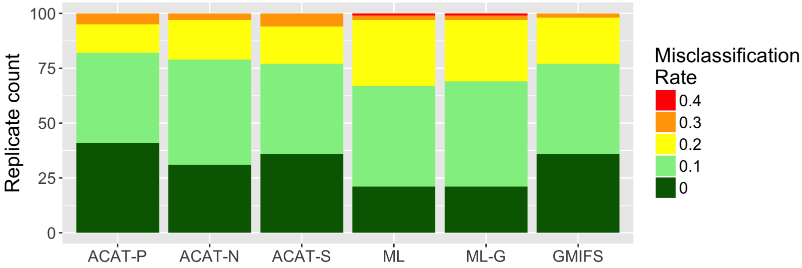
Our contribution

- ordinalNet: fits parallel/nonparallel/semi-parallel ELMO models with elastic net penalty

Method comparison: TCGA liver tissue data

- Compared methods for out-of-sample prediction accuracy.
- 100 cross validation replicates. Each replicate randomly split data into 46 training and 10 test observations.
- Compared 6 methods:
  - Adjacent category elastic net
    - Parallel
    - Nonparallel
    - Semi-parallel
  - Multinomial logistic regression elastic net (fit by **glmnet**)
    - Standard penalty
    - Grouped penalty
  - Adjacent category generalized monotone incremental forward stagewise algorithm (fit by **ordinalgmifs**)
- All elastic net models used $\alpha = 0.5$.
- All methods were tuned by 10-fold cross validation on each training fold to optimize out-of-sample log-likelihood.

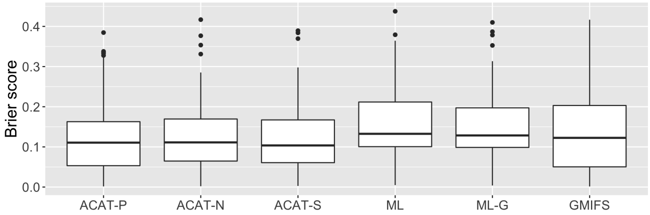
Method comparison: Misclassification rate



	Avg. misclassification rate
ACAT-Parallel	0.082
ACAT-Nonparallel	0.093
ACAT-Semiparallel	0.093
Multinomial Logistic	0.116
Multinomial Logistic-Grouped	0.114
GMIFS	0.089

Method comparison: Brier score

Note: Brier score is like mean squared error for categorical data (lower is better)



	Avg. Brier score
ACAT-Parallel	0.122
ACAT-Nonparallel	0.128
ACAT-Semiparallel	0.123
Multinomial Logistic	0.155
Multinomial Logistic-Grouped	0.153
GMIFS	0.132

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\hat{P}(Y_i = k|X_i = x_i) - I(y_i = k)\right)^2$$

Applications

ordinalNet has been cited in papers in a variety of applied fields:

- Respiratory disease (Albu, 2019)
- Major depressive disorder (Harati et al., 2019)
- Corporate bond rating (Park et al., 2018)
- Education (Crues et al., 2018)

Summary

- ELMO is a class of categorical response models that have parallel, nonparallel, and semi-parallel forms.
- Semi-parallel models have the flexibility of an unordered multinomial model, but can be penalized toward an ordinal model.
- Optimization done by coordinate descent.
- Implemented in **ordinalNet** package on CRAN.

Some related work

- (high-dimensional) variable selection for ordinal models
- flexible models for ordinal response data (e.g. non-proportional odds, multivariate ordinal data data, ...)
- Jan Gertheiss and co-authors, Helmut Schmidt University

  Tutz, Gerhard, and Jan Gertheiss. "Regularized regression for categorical data." Statistical Modelling 16.3 (2016): 161-200.

  Ugba ER. serp: An R package for smoothing in ordinal regression. Journal of Open Source Software. 2021 Oct 27;6(66):3705. "functions for regularization across response categories in the non-proportional cumulative ordinal regression model."
- Hirk, Hornik, and Vana, Vienna University of Economics and Business (hello!)

  Hirk R, Hornik K, Vana L. Multivariate ordinal regression models: an analysis of corporate credit ratings. Statistical Methods & Applications. 2019 Sep;28(3):507-39.

  Hirk R, Hornik K, Vana L. mvord: an R package for fitting multivariate ordinal regression models. Journal of Statistical Software. 2020 Apr 18;93:1-41.

Some related work (continued)

- Paul Rathouz (ordinal outcomes, not regularization). Semi-parametric generalized linear model (SPGLM).

  Wurm MJ, Rathouz PJ. Semi-parametric generalized linear models with the gldrm package. The R journal. 2018 Jul;10(1):288.

  ENAR March 2022. "Comparative Performance of a Semi-parametric Generalized Linear Model in Selected Analysis Settings." (session: Regression models for ordinal response data).

- Kellie Archer, Ohio State University.

  Zhang, Y.; Archer, K.J. Bayesian penalized cumulative logit model for high-dimensional data with an ordinal response. Statistics in Medicine 2021, 40, 1453–1481.

  Archer, Kellie J., et al. "Ordinalbayes: Fitting Ordinal Bayesian Regression Models to High-Dimensional Data Using R." (2022).

References

Albu, E. (2019). *Ventilator-associated Events Prediction*. Ph.D. thesis, Ghent University.

Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G. and Gentry, A. E. (2014). ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer Informatics*, **13** 187.

Archer, K. J., Mas, V. R., Maluf, D. G. and Fisher, R. A. (2010). High-throughput assessment of CpG site methylation for distinguishing between HCV-cirrhosis and HCV-associated hepatocellular carcinoma. *Molecular Genetics and Genomics*, **283** 341–349.

Crues, R., Bosch, N., Perry, M., Angrave, L., Shaik, N. and Bhat, S. (2018). Refocusing the lens on engagement in moocs. In *Proceedings of the fifth annual ACM conference on learning at scale*. ACM, 11.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1** 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33** 1–22.

Goeman, J. J., Meijer, R. J. and Chaturvedi, N. (2017). *Penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model*. R package version 0.9-50.

Harati, S., Crowell, A., Huang, Y., Mayberg, H. and Nemati, S. (2019). Classifying depression severity in recovery from major depressive disorder via dynamic facial features. *IEEE journal of biomedical and health informatics*.

Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Park, H., Kang, J., Heo, S. and Yu, D. (2018). Comparative study of prediction models for corporate bond rating. *Korean Journal of Applied Statistics*, **31** 367–382.

Peterson, B., Harrell, F. E. and Petersont, B. (1990). Partial Proportional Odds Models for Ordinal Response Variables. *Journal of the Royal Statistical Society: Series C*, **39** 205–217.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. 4th ed. Springer, New York. URL http://www.stats.ox.ac.uk/pub/MASS4.

Wilhelm, M. S., Carter, E. M. and Hubert, J. J. (1998). Multivariate iteratively re-weighted least squares, with applications to dose-response data. *Environmetrics*, **9** 303–315.

Wurm, M. J., Rathouz, P. J. and Hanlon, B. M. (2017a). *ordinalNet: Penalized Ordinal Regression*. R package version 2.0, URL https://CRAN.R-project.org/package=ordinalNet.

Wurm, M. J., Rathouz, P. J. and Hanlon, B. M. (2017b). Regularized ordinal regression and the ordinalnet r package. *arXiv preprint arXiv:1706.05003*.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, **32** 1–34. URL http://www.jstatsoft.org/v32/i10/.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer, New York, USA.

Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *Journal of Royal Statistical Society: Series B (Methoodlogical)*, **58** 481–493.