CS3390 Assignment 1
Problem 2

Gautam Singh (CS21BTECH11018)
Jaswanth Beere (BM21BTECH11007)

Contents

This document summarises and demonstrates the methods described in [1] using the dataset in [2].

1 Summary

Data is said to be *ordinal* if it can be grouped into ordered categories. These categories can be thought of as intervals of a function of some underlying, unknown continuous random variable, where these intervals can be of any distance and non-continuous. Two models for performing regression on ordinal data, namely the *proportional odds* and *proportional hazards* model, are described in [1] along with their practical use cases. Suppose that the response variable belongs to k ordered categories with probabilities $\pi_j(\mathbf{x})$. In both cases, the models are of the form

$$\text{link}\left\{\gamma_j(\mathbf{x})\right\} = \theta_j - \boldsymbol{\beta}^\top \mathbf{x} \tag{1}$$

which is a general linear model where

1) $\gamma_j(\mathbf{x}) = \sum_{i=1}^{j} \pi_i(\mathbf{x})$ denotes the probability that the response variable falls in category at most j for inputs $\mathbf{x}$.
2) $\boldsymbol{\beta}$ and θ_j, $1 \le j < k$ are the parameters of the model to be estimated. Usually, θ_j are referred to as *cut points*.
3) link is a monotone function mapping $(0, 1)$ to $(-\infty, \infty)$. The link function should be selected based on ease of interpretation for the application.

Also studied in this paper are nonlinear models of the form

$$\text{link}\left\{\gamma_j(\mathbf{x_i})\right\} = \frac{\theta_j - \boldsymbol{\beta}^\top \mathbf{x_i}}{\tau_i} \tag{2}$$

which are useful for applications where the covariates $\mathbf{x}$ may not have the same "variance". Here, τ_i is called the *scale* of the ith row of a dataset, and is given by

$$\log \tau_i = \boldsymbol{\tau}^\top (\mathbf{x_i} - \bar{\mathbf{x}}) \tag{3}$$

where $\boldsymbol{\tau}$ is a vector of parameters of the model.
In both cases, it turns out that numerical methods of solving these models, in particular reweighted least squares, converges to the maximum likelihood estimate. The paper also contrasts the performance of ordinal regression with application-specific qualitative tests and alternative models for various healthcare-related applications, illustrating that models such as (1) are easy to compute and interpret.

2 Differences From Other Methods

It is tempting to consider the ordinal categories as classes and use a classification model. However, in cases where the categories are models of underlying continuous random variables and processes, it is better to use ordinal regression. Using classification methods is amenable only when the proportion of classes in the sample dataset and population dataset are equal, which is not always possible. Classification in such cases can lead to indicators with poor generalization performance.
On the other hand, regression is suitable for applications where the response variable is real-valued. Obviously, applying regression directly for ordinal data will not work, since the data is discrete and grouped into categories. Further, it is difficult to obtain information about the underlying model or

process which puts the response variable into these intervals. Therefore, ordinal regression makes use of a link function which makes the model amenable to the same tools as for linear regression, such as least squares and gradient descent.

3 Maximum Likelihood Parameter Estimation

Suppose that the categorical responses in a dataset of n inputs are distributed as $\{n_1, n_2, \ldots, n_k\}$. Define for $1 \leq j \leq k$,

$$R_j \triangleq \sum_{i=1}^{j} n_i, \; Z_j \triangleq \frac{R_j}{n}. \tag{4}$$

Using the definitions of γ_j described earlier, define

$$\phi_j \triangleq \log\left(\frac{\gamma_j}{\gamma_{j+1} - \gamma_j}\right) = \operatorname{logit}\left(\frac{\gamma_j}{\gamma_{j+1}}\right) \tag{5}$$

$$g\left(\phi_j\right) \triangleq \log\left(1 + \exp\left(\phi_j\right)\right) = \log\left(\frac{\gamma_{j+1}}{\gamma_{j+1} - \gamma_j}\right). \tag{6}$$

From (5), we have

$$\frac{\partial \phi_j}{\partial \gamma_j} = \left(\frac{\gamma_{j+1} - \gamma_j}{\gamma_j}\right) \frac{\left(\gamma_{j+1} - \gamma_j\right) - \left(-\gamma_j\right)}{\left(\gamma_{j+1} - \gamma_j\right)^2} \tag{7}$$

$$= \frac{\gamma_{j+1}}{\gamma_j\left(\gamma_{j+1} - \gamma_j\right)} \tag{8}$$

$$\frac{\partial \phi_j}{\partial \gamma_{j+1}} = \left(\frac{\gamma_{j+1} - \gamma_j}{\gamma_j}\right) \frac{-\gamma_j}{\left(\gamma_{j+1} - \gamma_j\right)^2} \tag{9}$$

$$= \frac{-1}{\gamma_{j+1} - \gamma_j} \tag{10}$$

$$\implies \frac{\partial \phi_j}{\partial \gamma_j} = -\frac{\gamma_{j+1}}{\gamma_j} \frac{\partial \phi_j}{\partial \gamma_j} \tag{11}$$

Noting that $\gamma_k = 1$, the likelihood can be written as

$$L = \prod_{j=1}^{k} \pi_j^{n_i} \tag{12}$$

$$= \gamma_1^{R_1} \prod_{j=1}^{k-1} \left(\gamma_{j+1} - \gamma_i\right)^{R_{j+1} - R_j} \tag{13}$$

$$= \gamma_1^{R_1} \prod_{j=1}^{k-1} \gamma_{j+1}^{R_{j+1}} \left(\frac{1}{\gamma_{j+1}^{R_j}}\right) \left(\frac{\gamma_{j+1} - \gamma_j}{\gamma_{j+1}}\right)^{R_{j+1} - R_j} \tag{14}$$

$$= \prod_{j=1}^{k-1} \left(\frac{\gamma_j}{\gamma_{j+1}}\right)^{R_j} \left(\frac{\gamma_{j+1} - \gamma_j}{\gamma_{j+1}}\right)^{R_{j+1} - R_j} \tag{15}$$

$$= \prod_{j=1}^{k-1} \left(\frac{\gamma_j}{\gamma_{j+1} - \gamma_j}\right)^{R_j} \left(\frac{\gamma_{j+1} - \gamma_j}{\gamma_{j+1}}\right)^{R_{j+1}}. \tag{16}$$

Therefore, using (4), the log-likelihood is

$$l = n \sum_{i=1}^{k-1} \left(Z_i \phi_i - Z_{i+1} g\left(\phi_i\right)\right). \tag{17}$$

To find the maximum likelihood estimate of θ_j, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, [1] recommends the use of the Newton-Raphson Method with Fisher scoring. This method converges rapidly even for poor initial estimates, provided that the θ_j are monotone increasing.
Define

$$\boldsymbol{\beta}^* \triangleq \begin{pmatrix} \theta_1 & \theta_2 & \ldots & \theta_{k-1} & \beta_1 & \ldots & \beta_p \end{pmatrix}^\top \tag{18}$$

and denote the entire parameter vector as

$$\boldsymbol{\psi} \triangleq \begin{pmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\tau} \end{pmatrix}. \tag{19}$$

Define

$$\mathbf{X}_j^* \triangleq \begin{pmatrix} \mathbf{e}_j \\ \mathbf{x} \end{pmatrix} \tag{20}$$

$$\mathbf{U} \triangleq \bar{\mathbf{x}} - \mathbf{x} \tag{21}$$

$$w \triangleq \exp\left(\boldsymbol{\tau}^\top \mathbf{U}\right). \tag{22}$$

Then, (2) can be rewritten as

$$Y_j = \operatorname{link}\left(\gamma_j\right) = \boldsymbol{\beta}^{*\top} \mathbf{X}_j^* \exp\left(\boldsymbol{\tau}^\top \mathbf{U}\right). \tag{23}$$

Here, $\mathbf{e}_j^\top$ denotes the jth standard basic vector in $\mathbb{R}^{k-1}$.
From (23), we have

$$\frac{\partial Y_j}{\partial \beta_r^*} = X_{jr}^* \exp\left(\boldsymbol{\tau}^\top \mathbf{U}\right) = w X_{jr}^* \tag{24}$$

$$\frac{\partial Y_j}{\partial \tau_r} = U_r \left(\boldsymbol{\beta}^{*\top} \mathbf{X}_j^* \exp\left(\boldsymbol{\tau}^\top \mathbf{U}\right)\right) = Y_j U_r \tag{25}$$

Using the chain rule,

$$\frac{\partial l}{\partial \beta_r^*} = \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} \frac{\partial \phi_j}{\partial \beta_r^*} \tag{26}$$

$$= \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} \left(\frac{\partial \phi_j}{\partial \gamma_j} \frac{\partial \gamma_j}{\partial \beta_r^*} + \frac{\partial \phi_j}{\partial \gamma_{j+1}} \frac{\partial \gamma_{j+1}}{\partial \beta_r^*} \right) \tag{27}$$

$$= \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} \left(\frac{\partial \phi_j}{\partial \gamma_j} \frac{\partial \gamma_j}{\partial Y_j} \frac{\partial Y_j}{\partial \beta_r^*} + \frac{\partial \phi_j}{\partial \gamma_{j+1}} \frac{\partial \gamma_{j+1}}{\partial Y_{j+1}} \frac{\partial Y_{j+1}}{\partial \beta_r^*} \right) \tag{28}$$

$$= w \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} \frac{\partial \phi_j}{\partial \gamma_j} \left(\frac{\partial \gamma_j}{\partial Y_j} X_{jr}^* - \frac{\gamma_j}{\gamma_{j+1}} \frac{\partial \gamma_{j+1}}{\partial Y_{j+1}} X_{j+1,r}^* \right) \tag{29}$$

$$= w \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} V_j^{-1} q_{jr} \tag{30}$$

where

$$V_j \triangleq \frac{\partial \gamma_j}{\partial \phi_j} \tag{31}$$

$$q_{jr} \triangleq \frac{\partial \gamma_j}{\partial Y_j} X_{jr}^* - \frac{\gamma_j}{\gamma_{j+1}} \frac{\partial \gamma_{j+1}}{\partial Y_{j+1}} X_{j+1,r}^*. \tag{32}$$

From (5), (6) and (17), we have,

$$\frac{\partial l}{\partial \phi_j} = n \left(Z_j - Z_{j+1} \frac{\partial g\left(\phi_j\right)}{\partial \phi_j} \right) \tag{33}$$

$$= n \left(Z_j - Z_{j+1} \frac{\gamma_j}{\gamma_{j+1}} \right). \tag{34}$$

Thus, using (24) and (32),

$$\frac{\partial^2 l}{\partial \beta_r^* \partial \phi_j} = -nw Z_{j+1} \frac{\gamma_{j+1} \frac{\partial \gamma_j}{\partial Y_j} X_{jr}^* - \gamma_j \frac{\partial \gamma_{j+1}}{\partial Y_{j+1}} X_{j+1,r}^*}{\gamma_{j+1}^2} \tag{35}$$

$$= -nw \frac{Z_{j+1}}{\gamma_{j+1}} q_{jr}. \tag{36}$$

Notice that V_j and q_{jr} are independent of β_s^*, $s \neq r$. Differentiating l twice, from (30) and (36),

$$\frac{\partial^2 l}{\partial \beta_r^* \partial \beta_s^*} = -nw^2 \sum_{j=1}^{k-1} V_j^{-1} \frac{Z_{j+1}}{\gamma_{j+1}} q_{jr} q_{js} \tag{37}$$

$$\implies E\left[\frac{\partial^2 l}{\partial \beta_r^* \partial \beta_s^*} \right] = -nw^2 \sum_{j=1}^{k-1} V_j^{-1} q_{jr} q_{js} \tag{38}$$

since by (4), $E\left[Z_{j+1}\right] = \gamma_{j+1}$.

Using (25), (11) and the chain rule in a similar manner, we obtain

$$\frac{\partial l}{\partial \tau_r} = U_r \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} \frac{\partial \phi_j}{\partial \gamma_j} \left(\frac{\partial \gamma_j}{\partial Y_j} Y_j - \frac{\gamma_j}{\gamma_{j+1}} \frac{\partial \gamma_{j+1}}{\partial Y_{j+1}} Y_{j+1} \right) \tag{39}$$

$$= U_r \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} V_j^{-1} q_j \tag{40}$$

where we define

$$q_j \triangleq \frac{\partial \gamma_j}{\partial Y_j} Y_j - \frac{\gamma_j}{\gamma_{j+1}} \frac{\partial \gamma_{j+1}}{\partial Y_{j+1}} Y_{j+1}. \tag{41}$$

Thus, the expected second derivative turns out to be

$$E\left[\frac{\partial^2 l}{\partial \tau_r \partial \tau_s} \right] = -n U_r U_s \sum_{j=1}^{k-1} V_j^{-1} q_j^2. \tag{42}$$

Similarly, the mixed expected second derivatives are

$$E\left[\frac{\partial^2 l}{\partial \beta_r^* \partial \tau_s} \right] = -nw U_s \sum_{j=1}^{k-1} V_j^{-1} q_j q_{jr}. \tag{43}$$

Thus, the negative expectation of the Hessian matrix $\mathbf{A}_l = -E\left[\mathbf{H}_l\right]$ is symmetric and from (38) and (42), has nonnegative diagonal entries. Therefore, it is positive semidefinite.

Applying a Taylor series expansion for $\nabla_{\boldsymbol{\psi}} l$ around $\boldsymbol{\psi}_n$ using the expected Hessian matrix, we get

$$\nabla l\left(\boldsymbol{\psi}\right) = \nabla l\left(\boldsymbol{\psi}_n\right) - \mathbf{A}_l \left(\boldsymbol{\psi} - \boldsymbol{\psi}_n\right) + \ldots \tag{44}$$

Using a linear approximation, and taking $\nabla l\left(\boldsymbol{\psi}\right) = \mathbf{0}$ at $\boldsymbol{\psi} = \boldsymbol{\psi}_{n+1}$, we get the update equation

$$\mathbf{A}_l \left(\boldsymbol{\psi}_{n+1} - \boldsymbol{\psi}_n\right) = \nabla l\left(\boldsymbol{\psi}_n\right) \tag{45}$$

$$\implies \boldsymbol{\psi}_{n+1} = \boldsymbol{\psi}_n + \mathbf{A}_l^{-1} \nabla l\left(\boldsymbol{\psi}_n\right). \tag{46}$$

where the entries of $\mathbf{A}_l$ are defined in (38), (42) and (43).

References

[1] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980, Accessed Oct. 2, 2023. [Online]. Available: http://www.jstor.org/stable/2984952

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Wine Quality," UCI Machine Learning Repository, 2009, Accessed: Oct. 2, 2023. DOI: https://doi.org/10.24432/C56S3T.