FOML Assignment1

SADINENI ABHINAY - CS21BTECH11055
NIMMALA AVINASH - CS21BTECH11039
Group 64

October 2023

Question 2

(a) **Brief Summary of the research paper:**

- At first the document discuss the nature of ordinal data
- Two models odd-proportional and odds-hazard models are discussed as they are simple for interpretation
- These two models are shown to be multivariate extensions of generalized linear models.
- Paper next discuss the properties of related linear models three important properties are discussed
- Linear models for ordinal data typically have a general form where a link function is applied to the cumulative probabilities
- Linear models should ideally exhibit some form of in-variance under a reversal of category order for ordinal data analysis. While models like the logistic and probit are invariant under category order reversal, models like the complementary log-log may not be.
- When testing hypotheses in linear models for ordinal data, similar rank tests are often employed. These tests are based on the conditional distribution of the data given the marginal totals for the response
- the paper discuss the complex co-variate structure and parameter estimation of general model.
- In last two sections alternative models and analysis of residuals are discussed

Likelihood and odd ratio:

(i) **Likelihood**:(K classes or categories)

$$L = \prod_{i=1}^{n} P\left(Y_i = y_i\right) = \prod_{i=1}^{n} \left(\prod_{j=1}^{K} P\left(Y_i = j\right)^{\delta_{j,y_i}} \right) \tag{1}$$

- For both structure of likelihood function is same but difference occurs in variables that we are using in it. classes/categories are ordered in ordinal regression and classes are independent in multi-class classification.
- We often use cumulative probabilities in ordinal regression to derive the likelihood as the categories are ordered we model cumulative probabilities of categories . whereas for multi class the function **remains same**

$$P(Y_i = j) = P(Y_i \le j) - P(Y_i \le j-1) \tag{2}$$

$$L = \prod_{i=1}^{n} \left(\prod_{j=1}^{K} P\left(Y_i = j\right)^{\delta_{j,y_i}} \right) \tag{3}$$

$$= \prod_{i=1}^{n} \left(\prod_{j=1}^{K} \left(P(Y_i \le j) - P(Y_i \le j-1)\right)^{\delta_{j,y_i}} \right) \tag{4}$$

(ii) **Odd ratio**:

- Odds-Ratio is used in ordinal regression to to understand how the odds of moving from one category to another on the ordinal scale change as a predictor variable changes.For example take ordered logit model:

$$OR_{ij} = \frac{P(Y \le j)}{P(Y \le i)} \tag{5}$$

- In multi class classification the classes are independent the notion of odds-ratio do not typically apply here.

Difference between ordinal regression and regression problems:

- In regression , we typically use models like linear regression,polynomial regression to predict a continuous value. The goal is map the input features to a continuous output variable.
- In ordinal regression, we specialize models like ordered logit , ordered progit.These models are designed to estimate the probabilities of outcome falling into a ordinal category.
- likelihood of regression problem is different from that of ordinal regression . Usually we assume the errors are normally distributed

$$L = \prod_{i=1}^{n} P\left(Y_i = y_i\right) = \prod_{i} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(w^T.\phi(x_i)-t_i)^2} \tag{6}$$

- The notion of odds-ratio is not relevant here because the output is a continuous variable.

(b) **Parameter Estimation for ordinal data :**
Lets use derive ML estimate for general model ,for single multinomial observation $(x_1, x_2, ...x_n)$ expression for likelihood is given by product of k-1 quantites

$$L = \left\{ \left(\frac{\gamma_1}{\gamma_2}\right)^{\boldsymbol{R}_1} \left(\frac{\gamma_2 - \gamma_1}{\gamma_2}\right)^{\boldsymbol{R}_2 - \boldsymbol{R}_1} \right\} \left\{ \left(\frac{\gamma_2}{\gamma_3}\right)^{\boldsymbol{R}_2} \left(\frac{\gamma_3 - \gamma_2}{\gamma_3}\right)^{\boldsymbol{R}_3 - \boldsymbol{R}_2} \right\} \cdots \left\{ \left(\frac{\gamma_{k-1}}{\gamma_k}\right)^{\boldsymbol{R}_{k-1}} \left(\frac{\gamma_k - \gamma_{k-1}}{\gamma_k}\right)^{\boldsymbol{R}_k - \boldsymbol{R}_{k-1}} \right\}.$$

Notations:

- γ_i represents cumulative probabilities up to category i
- R_i represents cumulative counts up to category i $R_i = \sum_{j=1}^{i} n_j$
- $Z_i = \frac{R_i}{R_n}$

Lets us take proportional odds model

$$\gamma_j(x) = \frac{\exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}\right)}{1 + \exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}\right)}$$

$$\gamma_{j+1}(x) = \gamma_j(x) + (1 - \gamma_j(x))\gamma_{j+1}$$

(i) first expand the likelihood function

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{i=1}^{N} \prod_{j=1}^{n-1} \left\{ \left(\frac{\exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)}{1 + \exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)} \right)^{R_{ij}} \left(1 - \frac{\exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)}{1 + \exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)} \right)^{R_{ij+1}} \right\}$$

$$\log L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{n-1} \left[R_{ij} \left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right) - R_{ij+1} \log\left(1 + \exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)\right)\right]$$

(ii) now find the gradient of log-likelihood function

$$\nabla \log L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{n-1} \left[R_{ij} \mathbf{x}_i - R_{ij+1} \frac{\exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)}{1 + \exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)} \mathbf{x}_i \right]$$

(iii) construct Hessian matrix

$$\mathbf{H} = -\sum_{i=1}^{N} \sum_{j=1}^{n-1} R_{ij+1} \frac{\exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)}{\left(1 + \exp\left(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i\right)\right)^2} \mathbf{x}_i \mathbf{x}_i^T$$

(iv) use the hessian matrix for netwon ralpson method

$$\beta := \beta + H^{-1} \nabla \log L(\boldsymbol{\theta}, \boldsymbol{\beta})$$

Since we are not getting any closed form expression for parameter by setting it to zero we can use other methods for the estimation like netwon raphlson method,...Similarly find thresholds θ

(c) (Corresponding notebook file 2c.ipynb)

(a) **Linear Regression, LogisticIT(Ordinal regression) and multiclass logistic Regression** are the used models to predict the purity of the wine.

(b) Evalution metrics used to compare between linear regression and ordinal regression models are **MSE** and **MAE**, and between ordinal regression and multiclass logistic regression models are accuracy-score.

(c) MSE of linear regression is smaller compared to MSE of ordinal regression, MSE of linear regression is smaller than MAE of linear regression and MSE of ordinal regression is higher than MAE of ordinal regression because the MSE in linear regression squares the error value which is less than 1 which makes the value much lesser while in ordinal regression error values are integers and their squares are higher comparetively.

(d) File 2c.ipynb contains the implementation of predicting the purity of the wine.Intially the provided data is divided into two datasets X_train,y_train and X_test,y_test to train and test the model.And after training of the models test datasets are used to measure the MSE and MAE for both the models. (Cell 8 and Cell 9 contains the evalution metrics for white wine dataset similarly Cell 16 and Cell 17 contains evaluation metrics for red wine in 2c.ipynb).