Assignment - 1

Challa Akshay Santoshi, CS21BTECH11012
Cheekatla Hema Sri, CS21BTECH11013

October 8, 2023

Question 2

A multiclass classification problem with an order to the classes is known as an ordinal regression problem.

a

BRIEF SUMMARY

Various models utilize the ordinal nature of the data and share the property that the categories can be thought of as contiguous intervals on some contiguous scale. This can be done by describing various modes of stochastic ordering which eliminates the need for assigning scores or assuming cardinality instead of ordinality. They differ in their assumptions concerning the distributions of the latent variable like normality, homoscedasticity. The proportions-odd model is discussed and the logarithm of the odds function has a linear form which can be generalised with other linear models. For this, the link function, which is a transformation of the cumulative probabilities of the dependent ordered variable allows for the estimation of the model. The proportional hazards model is discussed. In this model, to obtain the appropriate linear structure analogous to the linear logistic model, the complementary log-log transform is used. The proportional odds and the proportional hazards model have the same general form called "link" which is a monotone function. These linear models are the multivariate extensions of the generalised linear models. Parameter estimation in general (including the non-linear models) introduces a new scale value for each row. It is shown that for the extension to non-linear models, the method of iteratively reweighted least squares converges to the maximum likelihood estimate, a property which greatly simplifies the necessary computation.

Odds ratio for ordinal regression:

Suppose that there are k ordered categories having probabilities $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), ..., \pi_k(\mathbf{x})$ when the covariates have the value $\mathbf{x}$. Let $\mathbf{Y}$ be the response and $k_j(\mathbf{x})$ denotes the odds that $\mathbf{Y} \leq j$, then the proportional odds model specifies that

$$k_j(\mathbf{x}) = k_j exp(-\beta^{\mathbf{T}}\mathbf{x}) \tag{1}$$

Here $(1 \leq j < k)$ and β is a vector of unknown parameters. The ratio of corresponding odds

$$k_j(\mathbf{x_1})/k_j(\mathbf{x_2}) = exp\{\beta^T(x_2 - x_1)\} \tag{2}$$

This expression is independent of j and depends only on the difference between the covariate values, $\mathbf{x_2} - \mathbf{x_1}$.

Odds ratio for multi-class classification:

For multi-class classification, the posterior probabilities given feature variables is given by

$$p(C_k|\phi) = y_k(\phi) = \frac{exp(a_k)}{\sum_j exp(a_j)} \tag{3}$$

where the 'activations' are given by

$$a_k = {\mathbf{w}_k}^T \phi \tag{4}$$

The odds for a given feature value ϕ_1 can be written as

$$\frac{p(C_k|\phi_1)}{1-p(C_k|\phi_1)} = \frac{exp({\mathbf{w}_k}^T\phi_1)}{\sum_j exp({\mathbf{w}_j}^T\phi_1) - exp({\mathbf{w}_k}^T\phi_1)} \tag{5}$$

The odds ratio for $\phi_\mathbf{1}$ and $\phi_\mathbf{2}$ would then be equal to

$$k_j(\phi_\mathbf{1})/k_j(\phi_\mathbf{2}) = \frac{exp({\mathbf{w}_j}^T\phi_1)}{\sum_k exp({\mathbf{w}_k}^T\phi_1) - exp({\mathbf{w}_j}^T\phi_1)} \times \frac{\sum_k exp({\mathbf{w}_k}^T\phi_2) - exp({\mathbf{w}_j}^T\phi_2)}{exp({\mathbf{w}_j}^T\phi_2)} \tag{6}$$

From equation (6), we can see that the odds ratio is dependent of j and further simplification is not possible. It cannot be reduced to the form of equation(2), therefore the odds ratio for ordinal regression and multi-class regression are different.

Likelihood for ordinal regression:

From equation (1), we can say that the odds for the event $\mathbf{Y} \leq j$, is the ratio $\gamma_j(\mathbf{x})/\{1-\gamma_j(\mathbf{x})\}$ where $\gamma_j(\mathbf{x}) = \pi_1(\mathbf{x}) + ... + \pi_j(\mathbf{x})$. This is identical to the linear logistic model given by

$$log[\gamma_j(\mathbf{x})/\{1-\gamma_j(\mathbf{x})\}] = \theta_j - \beta^\mathbf{T}\mathbf{x} \tag{7}$$

with $\theta_j = \log\, k_j$.
The contribution from a single multinomial observation $(n_1, ..., n_k)$ to the likelihood function is ${\pi_1}^{n_1}...{\pi_k}^{n_k}$.
Since we are dealing with cumulative probabilities, we define

$$\begin{aligned} R_1 &= n_1, \\ R_2 &= n_1 + n_2, \\ &\vdots \\ R_k &= \sum n_j = n \end{aligned}$$

The cell counts be $\{n_{ij}\}$ with row totals n1, n2 and column or category totals $\{n_{.j}\}$. The cumulative row sums are R_{ij} so that $n_i = R_{ik}$ is the ith row total.
In terms of the parameters of the cumulative transformation, the likelihood can be written as the product of k-1 quantities

$$\left\{ \left(\frac{\gamma_1}{\gamma_2}\right)^{R_1} \left(\frac{\gamma_2-\gamma_1}{\gamma_2}\right)^{R_2-R_1} \right\} \left\{ \left(\frac{\gamma_2}{\gamma_3}\right)^{R_2} \left(\frac{\gamma_3-\gamma_2}{\gamma_3}\right)^{R_3-R_2} \right\} \dots \left\{ \left(\frac{\gamma_{k-1}}{\gamma_k}\right)^{R_{k-1}} \left(\frac{\gamma_k-\gamma_{k-1}}{\gamma_k}\right)^{R_k-R_{k-1}} \right\} \tag{8}$$

These factors are respectively the probability given R_2 that the first two cells divide in the ratio R_1 $:R_2 - R_1$; the probability given R_3 that the proportion in cell 3 relative to cells 1 and 2 combined is $R_3 - R_2$ $:R_2$ and so on for the other components.

Likelihood for multi-class classification:

For the likelihood function we use the 1-of-K coding scheme in which the target vector t_n for a feature vector ϕ_n belonging to class C_k is a binary vector with all elements zero except for element k, which equals one. The likelihood function is then given by

$$p(\mathbf{T}|\mathbf{w}_1, ..., \mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}} \tag{9}$$

where $y_{nk} = y_k(\phi_n)$, and $\mathbf{T}$ is an $N \times K$ matrix of target variables with elements t_{nk}.

From equation(8) and equation(9), we can see that likelihood functions for ordinal regression and multi-class classification are different.

Comparison with regression problems:

In particular when there are only two response categories, equation (7) is equivalent to a log-linear model. In general, however, when the number of categories exceeds 2, the linear logistic model does not correspond to a log-linear structure. All linear models of this form are qualitatively similar and for any given data set the fits are often indistinguishable.

b

Parameter estimates can be obtained by iteratively reweighted least squares. From general form of equation (7), if we relax the assumption of constant "variance" or scale parameter on that continuum, we introduce the multiplicative model given by

$$link(\gamma_{ij}) = (\theta_j - \beta^{\mathbf{T}}\mathbf{x_i})/\tau_i \tag{10}$$

where the quantity $\beta^T x_i$ is called the "location" for the ith row and τ_i is called the "scale" for the ith row. This model is saturated in scale parameters since we have one scale parameter associated with each row of the table. If $\dim(\beta) \geq t - 1$ where t is the number of rows, we say that it is also saturated in location parameters.
Unsaturated scale models satisfy

$$log\tau_i = \tau^{\mathbf{T}}(\mathbf{x_i} - \bar{\mathbf{x}}) \tag{11}$$

where τ is the vector of unknown parameters to be estimated. Furthermore, the estimates of $\log(\tau_i)$ are likely to be more symmetrically distributed than those of τ_i, a property which greatly improves approximation based on normality.

The likelihood which we obtained, as shown in the equation (8) can be used.

The following technique can be used to get the parameter value β:

We can take the logarithm of equation (8) and take the derivative of this expression with respect to β. The γ values can be replaced with the corresponding π_i values which in turn can be written in terms of $\beta^{\mathbf{T}}\mathbf{x}$. This expression when equated to 0, gives the parameter β.

The general non-linear model can be written as

$$Y_j = link(\gamma_j) = \beta^{*\mathbf{T}}\mathbf{X_j}^{*}\mathbf{exp}(\tau^{\mathbf{T}}\mathbf{U}) \tag{12}$$

where $\beta^* = (\theta_1, \theta_2, ..., \theta_{k-1}, \beta_1, ..., \beta_p)$ is the vector of parameters in the location model. $X_j{}^* = (0,...,1,...,0,\mathbf{X})$, where the 1 occurs in position j and $\mathbf{U_k}$, the vector of parameters in the scale model is normalized so that $\sum_i \mathbf{U_i} = 0$. Let $\psi^T = (\beta^{*T}, \tau^T)$ be the complete parameter vector and w = $\exp(\tau^T U)$).
The derivative of the log-likelihood with respect to β^* is

$$\frac{\partial l}{\partial \beta_r^*} = w \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} \left\{ \frac{\partial \phi_j}{\partial \gamma_j} \frac{d\gamma_j}{d\mathbf{Y}_j} \mathbf{X^*}_{jr} + \frac{\partial \phi_j}{\partial \gamma_{j+1}} \frac{d\gamma_{j+1}}{d\mathbf{Y}_{j+1}} \mathbf{X^*}_{j+1,r} \right\} \tag{13}$$

Substituting $\mathbf{V}_j = \frac{\partial \gamma_j}{\partial \phi_j}$ and $\frac{\partial \phi_j}{\partial \gamma_{j+1}} = (-\gamma_j/\gamma_{j+1})\mathbf{V^{-1}}_j$ we obtain

$$\frac{\partial l}{\partial \beta_r^*} = w \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} \mathbf{V^{-1}}_j q_{jr}, \tag{14}$$

where

$$q_{jr} = \left\{ \frac{d\gamma_j}{d\mathbf{Y}_j} \mathbf{X^*}_{jr} - \frac{\gamma_j}{\gamma_{j+1}} \frac{d\gamma_{j+1}}{d\mathbf{Y}_{j+1}} \mathbf{X^*}_{j+1,r} \right\} \tag{15}$$

Similarly the expected second derivative is

$$A_{rs} = -E\left(\frac{\partial^2 l}{\partial \beta^*{}_r \partial \beta^*{}_s} \right) = nw^2 \sum_j \mathbf{V}^{-1}{}_j q_{jr} q_{js}. \tag{16}$$

For the scale parameter τ the derivatives are

$$\frac{\partial l}{\partial \tau_r} = \sum \frac{\partial l}{\partial \phi_j} \left\{ \frac{\partial \phi_j}{\partial \gamma_j} \frac{d\gamma_j}{d\mathbf{Y}_j} \mathbf{Y_j U_r} + \frac{\partial \phi_j}{\partial \gamma_{j+1}} \frac{d\gamma_{j+1}}{d\mathbf{Y}_{j+1}} \mathbf{Y_{j+1} U_r} \right\}, \tag{17}$$

which reduce to

$$\frac{\partial l}{\partial \tau_r} = \mathbf{U_r} \sum_j \frac{\partial l}{\partial \phi_j} \mathbf{V}^{-1}{}_j q_j \tag{18}$$

The Taylor series expansion for $\partial l / \partial \psi$ gives

$$\frac{\partial l}{\partial \psi} = \left. \frac{\partial l}{\partial \psi} \right|_{\psi = \hat{\psi}} + \mathbf{A}(\partial \psi) + ..., \tag{19}$$

where $\partial \psi = \hat{\psi} - \psi$ and $\mathbf{A}$ is the negative expected matrix of second derivatives. Hence, updated values of ψ are obtained by iterating on the equation

$$\mathbf{A}\psi_{n+1} = \mathbf{A}\psi_n + \frac{\partial l}{\partial \psi} = b. \tag{20}$$

Writing $W =$log $w = -\tau^T \mathbf{U}$, we have

$$(\mathbf{A}\psi)_r = nw(1+W) \sum V^{-1}{}_j q_j q_{jr} \tag{21}$$

for r$\leq$dim(β^*). Similarly, for s>dim(β^*),

$$(\mathbf{A}\psi)_s = nw(1+W) U_s \sum V^{-1}{}_j q^2{}_j \tag{22}$$

The corresponding elements of b are given by

$$(b)_r = nw \sum_j V^{-1}{}_j q_{jr} \left\{ q_j(1+W) + Z_j - \frac{\gamma_j}{\gamma_{j+1}} Z_{j+1} \right\} \tag{23}$$

and

$$(b) = nU_s \sum_j V^{-1}{}_j q_j \left\{ q_j(1+W) + Z_j - \frac{\gamma_j}{\gamma_{j+1}} Z_{j+1} \right\} \tag{24}$$

Finally, all the above equations represent the contribution to the log-likelihood and its derivatives from a single multinomial observation and the total contribution is the sum of the individual contributions.