ADS research project, ordinal data

F. Meijvis, D. Hessen, C. Spitoni

February 10, 2025

1 Definition of ordinal regression

The dependent random variable Y has ordinal level of measurement. The number of possible values that Y can take is $m+1$. A realization of Y is denoted by y and $y \in \{0, 1, \ldots, m\}$. The conditional probability distribution of Y given a set of interval predictors $\mathbf{x} = (x_1, \ldots, x_k)'$ is $P(Y = y \,|\, \mathbf{x})$. The regression of Y on $\mathbf{x}$ is ordinal if and only if

$$P(Y \leq y \,|\, \vec{x}) = \sum_{t=0}^{y} P(Y = t \,|\, \vec{x}), \quad \text{for all } y, \tag{1}$$

is a monotonic function of $\mathbf{x}$. If higher values of y are associated with higher values of $\mathbf{x}$, then this function should be non-increasing, for all y. If $P(Y \leq y \,|\, \mathbf{x})$ is a monotonic function of x_i, then its derivative with respect to x_i is either non-negative or non-positive for all x_i.

2 Multinomial logistic regression

Let $z_s = 1$ if $y = s$ and $z_s = 0$ otherwise, for $s = 1, \ldots, m$, then the conditional probability distribution of Y given $\mathbf{x}$ can be written as

$$P(Y = y \,|\, \vec{x}) = P(Y = 0 \,|\, \vec{x}) \prod_{s=1}^{m} \left\{ \frac{P(Y = s \,|\, \vec{x})}{P(Y = 0 \,|\, \vec{x})} \right\}^{z_s}, \tag{2}$$

where

$$P(Y = 0 \,|\, \vec{x}) = \left\{ 1 + \sum_{s=1}^{m} \frac{P(Y = s \,|\, \vec{x})}{P(Y = 0 \,|\, \vec{x})} \right\}^{-1}. \tag{3}$$

In the multinomial logistic regression model,

$$\frac{P(Y = s \,|\, \vec{x})}{P(Y = 0 \,|\, \vec{x})} = \exp(\alpha_{0s} + \vec{\alpha}_s' \vec{x}), \text{ for } s = 1, \ldots, m, \tag{4}$$

where $\vec{\alpha}_s = (\alpha_{1s}, \ldots, \alpha_{ks})'$, so that

$$P(Y = y \,|\, \vec{x}) = \frac{\exp\left\{ \sum_{s=1}^{m} z_s (\alpha_{0s} + \vec{\alpha}_s' \vec{x}) \right\}}{1 + \sum_{s=1}^{m} \exp(\alpha_{0s} + \vec{\alpha}_s' \vec{x})}. \tag{5}$$

It follows that

$$
P(Y \le y \,|\, \vec{x}) =
\begin{cases}
\dfrac{1}{1+\sum\limits_{s=1}^{m} \exp(\alpha_{0s}+\vec{\alpha}_s'\vec{x})}, & \text{for } y = 0, \\[4ex]
\dfrac{1+\sum\limits_{s=1}^{y} \exp(\alpha_{0s}+\vec{\alpha}_s'\vec{x})}{1+\sum\limits_{s=1}^{m} \exp(\alpha_{0s}+\vec{\alpha}_s'\vec{x})}, & \text{for } y > 0.
\end{cases}
\tag{6}
$$

The multinomial logistic regression model has $(k+1)m$ parameters.

3   conditions for α

In order to have the multinomial regression model satisfy the definition of ordinal regression, we require $P(Y \le y \,|\, \vec{x})$ to be monotonic in $\vec{x}$. We will assume that higher values in $\vec{x}$ correspond to higher y. Furthermore, we treat both these statements are component wise.

For monotonicity, we require:

$$
\frac{\partial P(Y \le y|\vec{x})}{\partial x_i} \le 0
$$

For every i in every $\vec{x}$. For class $y = 0$, this ultimately gives:

$$
\sum_{s=1}^{m} \alpha_{is} \exp(\alpha_{0s} + \vec{\alpha}_s'\vec{x}) \ge 0
$$

This is satisfied if every $\alpha_{is} \ge 0$, but this may not be the strongest condition yet.

Note to self: I still want to test if it is possible to have 1 particular $\alpha \le 0$

For clas $y > 0$, we find:

$$
(1 + \sum_{s=1}^{m} \exp(\alpha_{0s} + \vec{\alpha}_s'\vec{x}))(\sum_{s=1}^{y} \alpha_{is}\exp(\alpha_{0s} + \vec{\alpha}_s'\vec{x}))
$$
$$
-(1 + \sum_{s=1}^{y} \exp(\alpha_{0s} + \vec{\alpha}_s'\vec{x}))(\sum_{s=1}^{m} \alpha_{is}\exp(\alpha_{0s} + \vec{\alpha}_s'\vec{x})) \le 0
$$

Which reduces back to the previous equation when $y = 0$, and is trivially satisfied when $y = m$.