Ordinal logistic regression

Part 1 (Model)

Assume our dependent variable is ordinal with k levels/categories $y_i \in \{1, 2, ..., k\}$ and our independent variables are real vectors $x_i \in \mathcal{R}^m$, for $i = 1..n$.

The ordinal regression model (or ordered logit) is defined as follows:

$$y_i | t, \beta, x_i \sim \text{Categorical}(p_i),$$

where β is the vector of size m of coefficients and t is a $k+1$ vector of thresholds 0-indexed for convenience $t_0 = -\infty < t_1 = 0 < t_2 < t_3 < \cdots < t_{k-1} < t_k = \infty$.

The size k probability vector of probabilities for each of the k categories for the $i-$th observation is defined (component-wise) as:

$$p_i(j) = F(t_j - u_i) - F(t_{j-1} - u_i), j = 1..k$$

where F is the CDF of the standard logistic distribution (or inverse logit) and $u_i = \beta^T x_i$.

Implement the (log-)likelihood of this model and an algorithm that fits this model using maximum likelihood estimation and is able to make predictions for new observations. For optimization you may use any third-party optimization library that allows for box-constraints. The algorithm should converge in reasonable time even with numerical gradients.

Practical considerations:

- The thresholds have to be ordered $t_1 < t_2 < \cdots < t_{k-1}$. This is a constraint that is not trivial to maintain during optimization. Instead, use the *stick breaking* parametrization $t_0 = -\infty$, $t_1 = 0$, $t_2 = t_1 + \Delta_1$, ..., $t_{k-1} = t_{k-2} + \Delta_{k-1}$, $t_k = \infty$. That is, let Δ_i be the parameters and derive t_i from them. A simple box constraint $\Delta_i > \epsilon = 0$ will suffice to keep the t_i ordered.
- In theory, $\epsilon = 0$, in practice, however, we might, with certain datasets and optimization algorithms run into problems where a Δ is so small that we get 0 probability and $-\infty$ likelihood. Starting values for Δ are also important (standardizing independent variables X makes this easier).
- Convergence issues may also arise if we have perfect separability (two or more categories can be perfectly fit). In such cases infinitely many thresholds are optimal. And, of course, if we have perfect colinearity in X, as is the case with any linear model.

Part 2 (Application)

We have a dataset that contains information about 250 students' response to the question *Overall, how would you rate this course?* for some Master's course (answers are 5-level ordinal ranging from very poor to very good). We are interested in the relationship between this variable and other available information, which includes age, sex, year of study (1st or 2nd) and grades (% scored on the exam) for the course in question and 7 other compulsory courses that the students took during their undergraduate studies. A grade of 50 in the course in question means that the student did not successfully pass the exam.

Apply the model from Part 1 to this dataset:

- Fit the model from Part 1 on this dataset. How you prepare the independent variables and if you include the intercept is up to you.
- Estimate the model's log-loss using k-fold cross-validation (choice of k is yours). As a baseline for comparison use the naive model that always predicts $p_i = (.15, .1, .05, .4, .3)$. Keep in mind that cross-validation estimates of loss, like any estimate, contain uncertainty - include a measure of uncertainty, such as standard error of log-loss. **Extra credit:** Include the Random Forests algorithm in the comparison (you may use a third-party library or your implementation from Homework 2) and briefly discuss the result.
- Interpret the model coefficients - which independent variables affect the response and how? Is there a sensible practical explanation? **Extra credit:** Again, regression coefficients, like any estimate, contain uncertainty. Include a measure of uncertainty, such as bootstrapped confidence intervals (less difficult) or intervals based on asymptotic normality of MLE (more difficult, refer to literature). Reinterpret the coefficients with this extra information. Note that in practice we should never interpret uncertain quantities without such extra information about the uncertainty.