Ordinal Classification

1 Introduction

When a variable is ordinal, its categories can be ranked from low to high, but the distances between adjacent categories are unknown.

For example, so-called Likert scales ask respondents whether they strongly agree, agree, have no opinion, disagree or strongly disagree with a statement. For an example, see figure 1. Ordinal variables are often treated as if they were measured on an interval scale. The dependent categories are numbered sequentially, and the linear regression model is used. This involves the implicit assumption that the intervals between adjacent categories are equal. For example, the distance between strongly agreeing and agreeing is assumed to be the same as the distance between agreeing and being neutral on a Likert scale. Likewise, averaging ordinal values (as is done in Caracal) should be frowned upon.

Another example of an ordinal scale is the classification into not relevant, relevant and highly relevant of search results to a query.

2 Alternative route to logistic regression

For compactness we switch to vector notation and write $\beta^\top \mathbf{x}$ instead of $\beta_0 + \sum_{i=1}^{p} \beta_i x_i$. In the new notation β is the column vector $(\beta_0, \beta_1, \ldots, \beta_p)^\top$ and $\mathbf{x}$ is the row vector $(1, x_1, \ldots, x_p)$. $\beta^\top \mathbf{x}$ is the dot product of these two vectors, which is $\beta_0 + \sum_{i=1}^{p} \beta_i x_i$.

The common path to logistic regression is to start with the observation is that linear regression

$$y = \beta^\top \mathbf{x} + \varepsilon, \qquad E[\varepsilon] = 0,$$

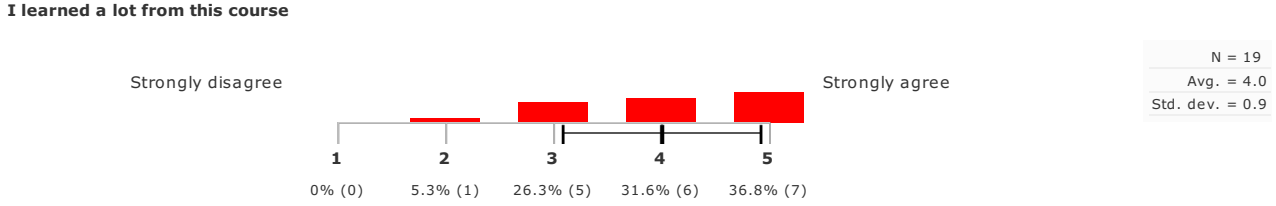

Figure 1: Excerpt from Caracal course evaluation

from which it follows that

$$E(y \mid \mathbf{x}) = P(y = 1 \mid \mathbf{x}) = \beta^\top \mathbf{x}$$

is not ideal for binary classification because the "probability estimates" produced are not constrained to lie between zero and one. They can be negative and they can be larger than one. To make sure they can't be negative, we can perform the transformation

$$P(y = 1 \mid \mathbf{x}) = e^{\beta^\top \mathbf{x}},$$

but now the probabilities can still be larger to one, so we take instead

$$P(y = 1 \mid \mathbf{x}) = \frac{e^{\beta^\top \mathbf{x}}}{1 + e^{\beta^\top \mathbf{x}}}.$$

We can arrive at the logistic regression and similar models via another path as well. We view the outcome $(y = 0, 1)$ as a discretization of an underlying regression. Consider for example the decision to make a large purchase. Micro-economic theory states that the consumer makes a cost-benefit calculation. Since benefit is not observable, we model the difference between cost and benefit as an unobserved variable y^*, such that

$$y^* = \beta^\top \mathbf{x} + \varepsilon, \qquad\qquad E[\varepsilon] = 0.$$

We do not observe the net benefit of the purchase, only whether it is made or not. Therefore, our observation is

$$y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases}$$

Now the probability that $y = 1$ is

$$\begin{aligned} P(y = 1 \mid \mathbf{x}) &= P(y^* > 0) \\ &= P(\beta^\top \mathbf{x} + \varepsilon > 0) \\ &= P(\varepsilon > -\beta^\top \mathbf{x}) \end{aligned}$$

If the distribution of ε is symmetric (e.g. normal or logistic), then

$$\begin{aligned} P(\varepsilon > -\beta^\top \mathbf{x}) &= P(\varepsilon < \beta^\top \mathbf{x}) \\ &= F(\beta^\top \mathbf{x}) \end{aligned}$$

Here F is the cumulative density function (cdf) of ε.

2.1 The probit and logit model

In the previous section we have established that

$$P(y = 1 \mid \mathbf{x}) = F(\beta^\top \mathbf{x})$$

In the so-called probit model we assume that ε has a standard normal distribution, that is, $\varepsilon \sim N(0,1)$. Thus, we have

$$P(y = 1 \mid \mathbf{x}) = \Phi(\beta^\top \mathbf{x})$$

where $\Phi(\cdot)$ is the standard normal cumulative density function.

The assumption of unit variance in the probit model is a harmless normalization. Suppose we assume that $\varepsilon \sim N(0, \sigma^2)$ as would be common in linear regression. We have

$$P(y = 1 \mid \mathbf{x}) = P(\varepsilon < \beta^\top \mathbf{x}) = P\left(\frac{\varepsilon}{\sigma} < \frac{\beta^\top \mathbf{x}}{\sigma}\right)$$

Now $\varepsilon/\sigma \sim N(0,1)$, so we can divide $\beta_0, \ldots, \beta_p$ by σ and get exactly the same probabilities as in the other model. Since we only observe whether y is 0 or 1 (and not the value of y^*), these models are observationally equivalent. The assumption of zero for the threshold is likewise innocent if the model contains a constant term β_0. An alternative parametrization is to fix $\beta_0 = 0$, i.e. we remove the intercept from the model:

$$y^* = \sum_{i=1}^{p} \beta_i x_i + \varepsilon.$$

Now we take

$$y = \begin{cases} 1 & \text{if } y^* > t \\ 0 & \text{if } y^* \leq t \end{cases}$$

In this case the threshold t has to be estimated from the data.

For the logit (logistic regression) model

$$P(y = 1 \mid \mathbf{x}) = \Lambda(\beta^\top \mathbf{x}) = \frac{e^{\beta^\top \mathbf{x}}}{1 + e^{\beta^\top \mathbf{x}}}$$

where $\Lambda(\cdot)$ indicates the logistic cumulative density function. See figure 2 for graphs of the logistic density and cumulative density functions. Note that the cumulative density function is our familiar logistic response function.

3 Ordinal Logit/Probit

Why did we go through all this trouble to present an alternative reasoning to arrive at the logistic regression model? Because we arrive at ordinal logistic regression via a similar route. Like before, we have a latent regression model

$$y_i^* = \beta^\top \mathbf{x}_i + \varepsilon_i, \qquad E[\varepsilon_i] = 0.$$

Again, y^* is not observed (latent variable), we only observe between which thresholds y^* falls. Let m denote the number of classes, where the classes are labeled $\{1, 2, \ldots, m\}$. Then

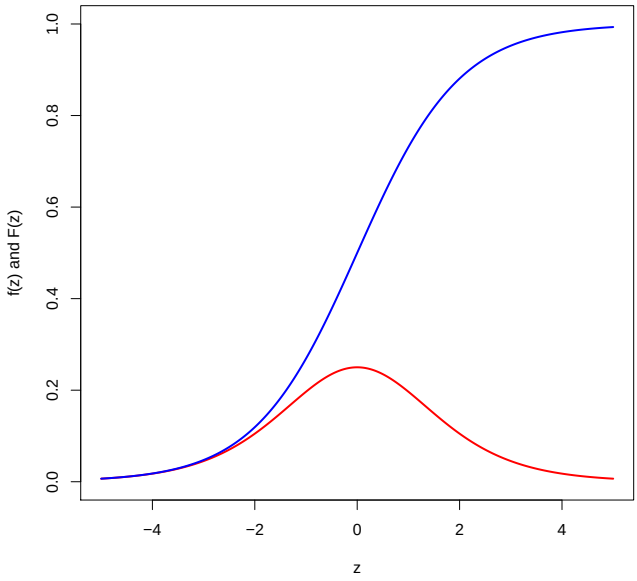

Figure 2: Logistic density (red) and cumulative density function (blue)

y is defined as follows:

$$y = \begin{cases} 1 & \text{if } -\infty < y^* \leq t_1 \\ 2 & \text{if } t_1 < y^* \leq t_2 \\ \vdots & \vdots \\ m & \text{if } t_{m-1} < y^* < \infty \end{cases}$$

Here $t_1, \ldots, t_{m-1}$ are unknown thresholds that have to be estimated from the data (together with the coefficient vector β). In this formulation the vector β does not contain an intercept β_0. Alternatively, we could include β_0 and fix one of the thresholds, e.g. we could fix t_1 to zero.

Let's first derive the formula for the probability that $y = 1$. We observe $y = 1$ when y^*

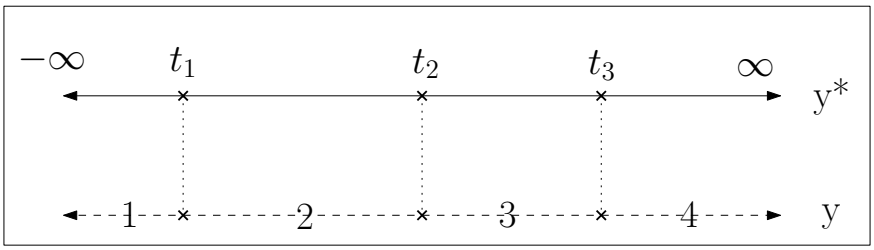

Figure 3: Relation between y^* (latent) and y (observed) for example with 4 class values.

falls between $t_0 = -\infty$ and t_1. Hence

$$P(y_i = 1 \mid \mathbf{x}_i) = P(t_0 \leq y_i^* < t_1 \mid \mathbf{x}_i)$$

Substituting $\beta^\top \mathbf{x}_i + \varepsilon_i$ for y_i^*, we get

$$P(y_i = 1 \mid \mathbf{x}_i) = P(t_0 \leq \beta^\top \mathbf{x}_i + \varepsilon_i < t_1 \mid \mathbf{x}_i)$$

Now we subtract $\beta^\top \mathbf{x}_i$ from all terms in the inequality to get

$$P(y_i = 1 \mid \mathbf{x}_i) = P(t_0 - \beta^\top \mathbf{x}_i \leq \varepsilon_i < t_1 - \beta^\top \mathbf{x}_i \mid \mathbf{x}_i)$$

Now recall that ε is a random variable with cumulative density function (cdf) denoted by F. The probability that a random variable is between two values is the difference between the cdf evaluated at these values. Therefore we have:

$$\begin{aligned} P(y_i = 1 \mid \mathbf{x}_i) &= P(\varepsilon_i < t_1 - \beta^\top \mathbf{x}_i \mid \mathbf{x}_i) - P(\varepsilon_i < t_0 - \beta^\top \mathbf{x}_i \mid \mathbf{x}_i) \\ &= F(t_1 - \beta^\top \mathbf{x}_i) - F(t_0 - \beta^\top \mathbf{x}_i) \end{aligned}$$

This derivation can be generalized to compute the probability of any observed outcome $y_i = j$ given $\mathbf{x}_i$. Thus we have the general rule that:

$$P(y_i = j \mid \mathbf{x}_i) = F(t_j - \beta^\top \mathbf{x}_i) - F(t_{j-1} - \beta^\top \mathbf{x}_i), \qquad j = 1, \ldots, m,$$

where $t_0 = -\infty$, and $t_m = \infty$.

When computing $P(y = 1 \,|\mathbf{x})$, the second term on the right hand side drops out, since $F(t_0 - \beta^\top \mathbf{x}_i) = F(-\infty - \beta^\top \mathbf{x}_i) = 0$. Likewise, when computing $P(y = m \,|\mathbf{x})$, the first term equals 1, since $F(t_m - \beta^\top \mathbf{x}_i) = F(\infty - \beta^\top \mathbf{x}_i) = 1$. So for a model with four possible classes, the formula's for the different outcomes are:

$$\begin{aligned} P(y_i = 1 \mid \mathbf{x}_i) &= F(t_1 - \beta^\top \mathbf{x}_i) \\ P(y_i = 2 \mid \mathbf{x}_i) &= F(t_2 - \beta^\top \mathbf{x}_i) - F(t_1 - \beta^\top \mathbf{x}_i) \\ P(y_i = 3 \mid \mathbf{x}_i) &= F(t_3 - \beta^\top \mathbf{x}_i) - F(t_2 - \beta^\top \mathbf{x}_i) \\ P(y_i = 4 \mid \mathbf{x}_i) &= 1 - F(t_3 - \beta^\top \mathbf{x}_i) \end{aligned}$$

Also, note that:

$$\begin{aligned} P(y_i \leq 1 \mid \mathbf{x}_i) &= F(t_1 - \beta^\top \mathbf{x}_i) \\ P(y_i \leq 2 \mid \mathbf{x}_i) &= F(t_2 - \beta^\top \mathbf{x}_i) \\ P(y_i \leq 3 \mid \mathbf{x}_i) &= F(t_3 - \beta^\top \mathbf{x}_i) \\ P(y_i \leq 4 \mid \mathbf{x}_i) &= 1 \end{aligned}$$

In general we have $P(y_i \leq j \mid \mathbf{x}_i) = F(t_j - \beta^\top \mathbf{x}_i)$.

3.1 Interpretation

The marginal effect of x_k is the slope of the curve relating x_k to $P(y = j \mid \mathbf{x})$, holding all other variables constant. Recall that

$$P(y = j \mid \mathbf{x}) = F(t_j - \beta^\top \mathbf{x}) - F(t_{j-1} - \beta^\top \mathbf{x}), \qquad j = 1, \ldots, m,$$

Taking the partial derivative with respect to x_k we get:

$$\begin{aligned}
\frac{\partial P(y = j \mid \mathbf{x})}{\partial x_k} &= \frac{\partial F(t_j - \beta^\top \mathbf{x})}{\partial x_k} - \frac{\partial F(t_{j-1} - \beta^\top \mathbf{x})}{\partial x_k} \\
&= \beta_k F'(t_{j-1} - \beta^\top \mathbf{x}) - \beta_k F'(t_j - \beta^\top \mathbf{x}) \\
&= \beta_k \left[f(t_{j-1} - \beta^\top \mathbf{x}) - f(t_j - \beta^\top \mathbf{x}) \right],
\end{aligned}$$

where $F' = f$, that is f is the pdf corresponding to the cdf F.

The sign of the marginal effect of x_k is not necessarily the same as the sign of β_k, since $f(t_{j-1} - \beta^\top \mathbf{x}) - f(t_j - \beta^\top \mathbf{x})$ can be negative, and even change sign.

Since the marginal effect of x_k depends on the value at which we hold the other variables constant, as well as on the value of x_k itself, we must decide on which values of the variables to use when computing the effect. One possibility is to compute the average marginal effect over all observations in the training sample:

$$\text{mean} \frac{\partial P(y = j \mid \mathbf{x})}{\partial x_k} = \frac{1}{n} \sum_{i=1}^{n} \beta_k \left[f(t_{j-1} - \beta^\top \mathbf{x}_i) - f(t_j - \beta^\top \mathbf{x}_i) \right]$$

In general the marginal effect does not indicate the change in the probability that would be observed for a unit change in x_k. However, if x_k varies over a region of the probability curve that is nearly linear, then the marginal effect can be used to summarize the effect of a unit change in x_k on the probability of an outcome.

It provides more insight to interpret the model in terms of cumulative probabilities of the class variable. Recall that

$$P(y_i \leq j \mid \mathbf{x}_i) = F(t_j - \beta^\top \mathbf{x}_i).$$

Hence

$$\begin{aligned}
\frac{\partial P(y \leq j \mid \mathbf{x})}{\partial x_k} &= \frac{\partial F(t_j - \beta^\top \mathbf{x})}{\partial x_k} \\
&= -\beta_k F'(t_j - \beta^\top \mathbf{x}) \\
&= -\beta_k f(t_j - \beta^\top \mathbf{x}).
\end{aligned}$$

Note that $f(t_j - \beta^\top \mathbf{x})$ is always positive, since f is a probability density function. So if β_k is positive, an increase in x_k will lead to a decrease in $P(y \leq j)$ for all $j = 1, \ldots, m-1$. In other words, an increase in x_k will lead to an increase in $P(y \geq j)$ for all $j = 2, \ldots, m$. In

this specific sense, one can say that if x_k increases, higher values of y become more likely. In my view this the easiest way to see the fundamental difference between an ordinal classification model and unordered one like the multinomial logit model. The ordinal model pre-supposes there is a *monotone* relationship between the predictor variables and the class variable. Depending of the sign of the coefficient β_k of the predictor x_k it can be summarized by the following "slogans":

- If β_k is positive: the higher the value of x_k, the more likely the higher class values.
- If β_k is negative: the higher the value of x_k, the more likely the lower class values.

4 Estimation

Recall that

$$P(y_i = j \mid \mathbf{x}_i) = F(t_j - \beta^\top \mathbf{x}_i) - F(t_{j-1} - \beta^\top \mathbf{x}_i), \qquad j = 1, \ldots, m,$$

where $t_m = \infty$ and $t_0 = -\infty$.

Hence, the likelihood function is

$$\begin{aligned} L(\beta, t \mid \mathbf{X}, \mathbf{y}) &= \prod_{j=1}^{m} \prod_{i:y_i=j} P(y_i = j \mid \mathbf{x}_i, \beta, t) \\ &= \prod_{j=1}^{m} \prod_{i:y_i=j} \left[F(t_j - \beta^\top \mathbf{x}_i) - F(t_{j-1} - \beta^\top \mathbf{x}_i) \right], \end{aligned}$$

where $\prod_{i:y_i=j}$ indicates we multiply over all cases where y is observed to have value j. Taking logs, we obtain the log likelihood function

$$\log L(\beta, t \mid \mathbf{X}, \mathbf{y}) = \sum_{j=1}^{m} \sum_{i:y_i=j} \log \left[F(t_j - \beta^\top \mathbf{x}_i) - F(t_{j-1} - \beta^\top \mathbf{x}_i) \right].$$

This expression can be maximized with numerical methods to estimate the t's and β's. We won't bother with the details.

5 Proportional Odds Logistic Regression

In the ordered logistic regression model we have:

$$\frac{P(y \leq j \mid \mathbf{x})}{P(y > j \mid \mathbf{x})} = \exp(t_j - \beta^\top \mathbf{x})$$

This can be seen as follows. Recall that

$$P(y \leq j \mid \mathbf{x}) = F(t_j - \beta^\top \mathbf{x}).$$

In logistic regression we choose for F the logistic cdf

$$\Lambda(z) = \frac{\exp(z)}{1+\exp(z)},$$

so we get

$$P(y \leq j \mid \mathbf{x}) = \frac{\exp(t_j - \beta^\top \mathbf{x})}{1+\exp(t_j - \beta^\top \mathbf{x})}.$$

Note also that

$$P(y > j \mid \mathbf{x}) = 1 - P(y \leq j \mid \mathbf{x}) = \frac{1}{1+\exp(t_j - \beta^\top \mathbf{x})},$$

since $P(y \leq j \mid \mathbf{x}) + P(y > j \mid \mathbf{x}) = 1$. Hence we get

$$\begin{aligned}\frac{P(y \leq j \mid \mathbf{x})}{P(y > j \mid \mathbf{x})} &= \frac{\exp(t_j - \beta^\top \mathbf{x})/1+\exp(t_j - \beta^\top \mathbf{x})}{1/1+\exp(t_j - \beta^\top \mathbf{x})} \\ &= \exp(t_j - \beta^\top \mathbf{x})\end{aligned}$$

The quantity

$$\Omega_j(\mathbf{x}) = \frac{P(y \leq j \mid \mathbf{x})}{P(y > j \mid \mathbf{x})}$$

is called the odds of the event $y \leq j$ against the event $y > j$ happening. To determine the effect of a change in $\mathbf{x}$, consider two values of $\mathbf{x}$, say $\mathbf{x} = \mathbf{x}'$ and $\mathbf{x} = \mathbf{x}^*$. The odds ratio at $\mathbf{x}'$ versus $\mathbf{x}^*$ equals

$$\frac{\Omega_j(\mathbf{x}')}{\Omega_j(\mathbf{x}^*)} = \frac{\exp(t_j - \beta^\top \mathbf{x}')}{\exp(t_j - \beta^\top \mathbf{x}^*)} = \exp\left(\beta^\top(\mathbf{x}^* - \mathbf{x}')\right),$$

where we used the rule that $\frac{e^x}{e^y} = e^{x-y}$. Notice that the odds ratio does not depend on the class j anymore. The odds depended on the class through t_j, but this term cancelled when we took the odds ratio. If x_k increases by 1, the odds ratio equals

$$\frac{\Omega_j(\mathbf{x}, x_k+1)}{\Omega_j(\mathbf{x}, x_k)} = \exp(-\beta_k)$$

To illustrate the interpretation using odds ratios consider the coefficient for gender (female=0, male=1) in the example in section 6. We have $\hat{\beta}_2 = -0.73$, so the odds ratio of male versus female is $\exp(-\hat{\beta}_2) = \exp(0.73) = 2.1$. This means that the odds of SD versus the combined outcomes D, A, and SA are 2.1 times greater for men than for women, holding all other variables equal (at any value). Likewise, the odds of SD and D versus A and SA are 2.1 times greater for men then for women, and finally the odds of SD, D and A versus SA are 2.1 times larger for men than for women.

This example illustrates that the odds ratio

$$\frac{\Omega_j(\mathbf{x}, x_k + 1)}{\Omega_j(\mathbf{x}, x_k)}$$

is the same for all values of j. This is known as the proportional odds assumption, and the reason ordinal logistic regression is often called proportional odds logistic regression. One can perform a test to determine whether this assumption is justified for a specific data set, but we will not discuss this. Remember: all models are wrong but some are more useful than others. (This slogan should however not be used as an excuse to not check the assumptions of your model!)

Finally, from

$$\frac{P(y \leq j \mid \mathbf{x})}{P(y > j \mid \mathbf{x})} = \exp(t_j - \beta^\top \mathbf{x}),$$

it follows that

$$\log\left[\frac{P(y \leq j \mid \mathbf{x})}{P(y > j \mid \mathbf{x})}\right] = t_j - \beta^\top \mathbf{x}.$$

Hence, we can view the proportional odds logistic regression model as a collection of parallel logistic regression models of $y \leq j$ against $y > j$. The fact that the decision boundaries for $y \leq j$ against $y > j$ run parallel to each other can be seen from the fact that the coefficient vectors are all the same, that is, we only have a single vector of coefficients β. The decision boundaries of $y \leq j$ against $y > j$ only differ in the threshold t_j. Also the decision boundaries of $P(y = j)$ against $P(y = k)$ run parallel to each other. This is again an important difference with the multinomial logit model, where the decision boundaries of $P(y = j)$ against $P(y = k)$ are linear but can have an arbitrary orientation, since each class j has its own vector of coefficients β_j.

6 Example Analysis in R

In 1977 and 1989, the General Social Survey asked respondents to evaluate the following statement:

> A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.

Responses were coded in the variable `warm` as: 1=Strongly Disagree (SD), 2=Disagree (D), 3=Agree (A), and 4=Strongly Agree (SA). The other variables are `yr89` (1 if survey year = 1989, 0 otherwise), `male` (1 if male, 0 if female), `white` (1 if white, 0 if nonwhite), `age` (age in years), `ed` (years of education), and `prst` (occupational prestige measured on some scale).

```
# load package containing the data
```

```
> library(Rchoice)
> data(Attitudes)
> dim(Attitudes)
[1] 2293    7

# convert the class label from integer to factor

> Attitudes[,1] <- as.factor(Attitudes[,1])

# load library containing proportional odds logistic regression

> library(MASS)

# fit proportional odds logistic regression (polr) model

> attitudes.polr <- polr(warm~.,data=Attitudes,Hess=T)

# show result; notice we have the following parameters: one vector of 6 coefficients
# and three thresholds (or intercepts)

> summary(attitudes.polr)
Call:
polr(formula = warm ~ ., data = Attitudes, Hess = T)

Coefficients:
          Value Std. Error t value
yr89   0.523912   0.079899   6.557
male  -0.733309   0.078483  -9.344
white -0.391140   0.118381  -3.304
age   -0.021666   0.002469  -8.777
ed     0.067176   0.015975   4.205
prst   0.006072   0.003293   1.844

Intercepts:
    Value    Std. Error t value
1|2  -2.4654   0.2389   -10.3188
2|3  -0.6309   0.2333    -2.7042
3|4   1.2618   0.2340     5.3919

Residual Deviance: 5689.825
AIC: 5707.825

# use fitted model to predict class labels on the training data
```

```
> attitudes.polr.pred <- predict(attitudes.polr,Attitudes,type="class")

# make table of true class labels against predictions
# notice that the model hardly predicts the extreme classes

> confmat.ord <- table(Attitudes[,1],attitudes.polr.pred)
> confmat.ord
   attitudes.polr.pred
      1   2   3   4
  1   7 162 128   0
  2   6 338 373   6
  3   2 208 624  22
  4   0  67 329  21

# compute percentage correctly predicted

> sum(diag(confmat.ord))/sum(confmat.ord)
[1] 0.4317488

# what is accuracy of predicting majority class?

> summary(Attitudes[,1])
  1   2   3   4
297 723 856 417

> 856/2293
[1] 0.3733101


> library(nnet)

# fit multinomial logit model on the same data

> attitudes.multinom <- multinom(warm~.,data=Attitudes)

# show results; notice we have the following parameters: 3 vectors of 7 coefficients

> summary(attitudes.multinom)
Call:
multinom(formula = warm ~ ., data = Attitudes)

Coefficients:
```

```
  (Intercept)      yr89       male      white          age         ed         prst
2    0.413324 0.7346215  0.1002630 -0.4215835 -0.002448876 0.09225126 -0.008866166
3    1.115388 1.0976382 -0.3597701 -0.5339769 -0.025004633 0.11056650  0.002433260
4    0.722171 1.1601947 -1.2264598 -0.8342253 -0.031676487 0.14357959  0.004165597

Std. Errors:
  (Intercept)      yr89      male     white         age         ed        prst
2   0.4290490 0.1656882 0.1410895 0.2472643 0.004424963 0.02734310 0.006157066
3   0.4303332 0.1636995 0.1411252 0.2463268 0.004482546 0.02803017 0.006138661
4   0.4928702 0.1810494 0.1676910 0.2641762 0.005218281 0.03377931 0.007002566

Residual Deviance: 5641.996
AIC: 5683.996

# use fitted model to predict class labels on the training data

> attitudes.multinom.pred <- predict(attitudes.multinom,Attitudes,type="class")

# make table of true class labels against predictions
# again the extreme classes are hardly ever predicted

> confmat.multinom <- table(Attitudes[,1],attitudes.multinom.pred)
> confmat.multinom
   attitudes.multinom.pred
      1   2   3   4
  1   4 162 131   0
  2   6 328 387   2
  3   1 219 630   6
  4   3  70 333  11

# compute percentage correctly predicted

> sum(diag(confmat.multinom))/sum(confmat.multinom)
[1] 0.4243349


# compare predictions of the two models

# make a vector indicating whether the multinom prediction is correct
> multinom.correct <- as.numeric(attitudes.multinom.pred == Attitudes[,1])
# make a vector indicating whether the polr prediction is correct
> polr.correct <- as.numeric(attitudes.polr.pred == Attitudes[,1])
```

```
# make a cross table of these two vectors
> table(polr.correct,multinom.correct)
            multinom.correct
polr.correct    0    1
           0 1218   85
           1  102  888
```

There are 85 cases that are predicted correctly by `multinom` and incorrectly by `polr`. There are 102 cases that are predicted correctly by `polr` and incorrectly by `multinom`. There is a total of 102+85=187 cases where one is wrong and the other is correct. If both classifiers would have the same accuracy, then for each of these 187 cases the probability would be 0.5 that `polr` wins and 0.5 that `multinom` wins. In other words, we would expect 93.5 cases in the (1,0) cell and 93.5 cases in the (0,1) cell. Let's call a win for `polr` a success. How probable is the observed number of successes (or a more extreme number) under the null hypothesis that both classifiers have the same accuracy? To compute this we must compute the probability of getting 102 successes or more, plus the probability of getting 85 successes or less.

```
# probability of observing 85 successes or less under the null hypothesis
> pbinom(85,187,prob=0.5)
[1] 0.1209552
# probability of observing 102 successes or more under the null hypothesis
> 1-pbinom(101,187,prob=0.5)
[1] 0.1209552
```

Hence the p-value is about 0.24, which means that the observed difference in accuracy is not unlikely under the null hypothesis that the true accuracies are the same. In other words, the observed difference in accuracy is not significant.