CS&SS/STAT/SOC 536: Regression for Ordinal Outcomes

Adrian Dobra
adobra@uw.edu

1 Introduction

We will assume that the outcome y takes a small number of values that have an intrinsic ordering. For example, y can take four values: 1 ("strongly disagree"), 2 ("disagree"), 3 ("agree") and 4 ("strongly agree"). The regression models for y will be a direct extension of the regression models for binary outcomes we discussed so far: if y takes only two values, it does not matter whether the first value is considered to be smaller than the second or viceversa. Therefore the probit and logit regressions for binary outcomes extend naturally to regressions for ordinal outcomes.

For simplicity we consider the case of a single explanatory (independent) variable x. We model the relationship between the latent variable y^* and the explanatory variable x using a simple linear regression:

$$y^* = \beta_0 + \beta_1 x + \epsilon.$$

The latent variable y^* is allowed to take any real value between $-\infty$ and $+\infty$. The relationship between y^* and y is determined by three parameters τ_1, τ_2, τ_3 called thresholds or cutpoints:

$$y = \begin{cases} 1, & \text{if } -\infty < y^* < \tau_1 \\ 2, & \text{if } \tau_1 \leq y^* < \tau_2 \\ 3, & \text{if } \tau_2 \leq y^* < \tau_3 \\ 4, & \text{if } \tau_3 \leq y^* < +\infty \end{cases}$$

Remark that we had only one threshold $\tau_1 = 0$ if y is binary. This should give you a hint that the model we defined before is not identifiable unless additional constraints are placed on the cutpoints τ_1, τ_2, τ_3.

We denote by f the density of the errors ϵ and by F the corresponding CDF of ϵ. In other words, we know that:

$$P(\epsilon \leq \epsilon_0) = F(\epsilon_0), \text{ and } \frac{dF(\epsilon_0)}{d\epsilon_0} = f(\epsilon_0).$$

The *ordered logit model* is obtained when the errors are assumed to follow the logistic distribution. The *ordered probit model* is obtained when the errors are assumed to be distributed

$N(0,1)$. [*Please see your handout on regression for binary outcomes for the definitions of the logistic and normal density and CDF*]

The probability of each value of y are given by:

$$\begin{aligned} P(y=1|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3) &= P(-\infty < y^* < \tau_1|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3), \\ &= P(-\infty < \beta_0+\beta_1 x+\epsilon < \tau_1), \\ &= P(\epsilon < \tau_1-\beta_0-\beta_1 x), \\ &= F(\tau_1-\beta_0-\beta_1 x). \end{aligned}$$

In a similar manner we obtain that:

$$\begin{aligned} P(y=2|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3) &= F(\tau_2-\beta_0-\beta_1 x)-F(\tau_1-\beta_0-\beta_1 x), \\ P(y=3|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3) &= F(\tau_3-\beta_0-\beta_1 x)-F(\tau_2-\beta_0-\beta_1 x), \\ P(y=4|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3) &= 1-F(\tau_3-\beta_0-\beta_1 x), \end{aligned}$$

These equations define the probability function of y given x (this is your regression model!). The same model can be specified by the distribution function of y given x:

$$\begin{aligned} P(y\leq 1|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3) &= F(\tau_1-\beta_0-\beta_1 x), \\ P(y\leq 2|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3) &= F(\tau_2-\beta_0-\beta_1 x), \\ P(y\leq 3|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3) &= F(\tau_3-\beta_0-\beta_1 x). \end{aligned}$$

Remark that $P(y\leq 4|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3)=1$, hence we do not need this fourth relationship to fully specify our model.

2 Making your model identifiable

The regression we have just defined involves five parameters: $(\beta_0,\beta_1,\tau_1,\tau_2,\tau_3)$. For any real number δ, consider the set of parameters $(\beta_0-\delta,\beta_1,\tau_1-\delta,\tau_2-\delta,\tau_3-\delta)$. For each $i=1,2,3$, we have:

$$F(\tau_i-\beta_0-\beta_1 x)=F((\tau_i-\delta)-(\beta_0-\delta)-\beta_1 x),$$

This implies

$$P(y\leq j|x,\beta_0,\beta_1,\tau_1,\tau_2,\tau_3)=P(y\leq j|x,\beta_0-\delta,\beta_1,\tau_1-\delta,\tau_2-\delta,\tau_3-\delta),$$

for each $j=1,2,3,4$. Therefore $(\beta_0-\delta,\beta_1,\tau_1-\delta,\tau_2-\delta,\tau_3-\delta)$ define the same regression model as $(\beta_0,\beta_1,\tau_1,\tau_2,\tau_3)$! In other words, you will not be able to produce a set of unique estimates of your regression coefficients unless you impose extra constraints on the regression parameters. For example, you could impose $\tau_1=0$ or $\beta_0=0$ (BUT NOT BOTH! This is why you cannot fit a logistic regression without an intercept β_0). Either one of these two

conditions will make your model identifiable, i.e. the estimates of your regression parameters will be unique. Make the connection with regression for binary outcomes: we imposed $\tau_1 = 0$ to make that model identifiable!

Warning: Various software packages make the model identifiable by imposing *different* constraints on the regression parameters. You should first try to understand what type of constraints each software package is using. DO NOT BE SURPRISED IF YOU GET DIFFERENT COEFFICIENT ESTIMATES WHEN FITTING YOUR MODEL WITH DIFFERENT LIBRARIES/STATISTICAL PACKAGES. This is fine as long as you understand what is going on.

3 Maximum Likelihood Estimation

We assume to have observed the independent samples $(y_1, x_1), \ldots, (y_n, x_n)$. We make the model identifiable by setting $\tau_1 = 0$ (alternatively, we could have chosen to set $\beta_0 = 0$). We omit τ_1 from the list of model parameters (since its value is now known) and write the likelihood as:

$$\begin{aligned} L(\beta_0, \beta_1, \tau_2, \tau_3) &= \prod_{j=1}^{4} \prod_{y_i = j} P(y_i = j | x_i, \beta_0, \beta_1, \tau_2, \tau_3), \\ &= \prod_{j=1}^{4} \prod_{y_i = j} \left[F(\tau_j - \beta_0 - \beta_1 x) - F(\tau_{j-1} - \beta_0 - \beta_1 x) \right]. \end{aligned}$$

In order to determine the MLES $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\tau}_2, \widehat{\tau}_3)$ we need to solve the system of equations:

$$\frac{\partial L}{\partial \beta_0} = 0, \quad \frac{\partial L}{\partial \beta_1} = 0, \quad \frac{\partial L}{\partial \tau_2} = 0, \quad \frac{\partial L}{\partial \tau_3} = 0.$$

This system does not admit an explicit solution and numerical methods (e.g., Newton-Raphson) are needed to solve it. From your point of view, you need to make sure your software did not give you any warnings about the convergence of its optimization procedure. If your MLEs do not look right or if you saw any warnings, you need to either increase the number of iterations in the optimization procedure or make sure your data has enough samples, does not contain coding errors, etc. *Do not use or report your MLEs unless you managed to eliminate all the convergence issues.*

After you have successfully determined the MLEs, your fitted model can be summarized by the predicted probabilities

$$P(y = j | x, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\tau}_2, \widehat{\tau}_3) = F(\widehat{\tau}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x) - F(\widehat{\tau}_{j-1} - \widehat{\beta}_0 - \widehat{\beta}_1 x), \quad j = 1, 2, 3, 4,$$

or by the cumulative probabilities

$$P(y \leq j | x, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\tau}_2, \widehat{\tau}_3) = F(\widehat{\tau}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x), \quad j = 1, 2, 3.$$

Remember that $\widehat{\tau}_1 = 0$.

4 Interpreting the regression parameters

The most straightforward way to interpret the slope β_1 is to make use of the latent variable y^* since its relationship with the explanatory variable x is linear:

$$\frac{\partial y^*(x)}{\partial x} = \widehat{\beta}_1.$$

That is, a unit change in x corresponds with a change of β_1 in y^*. Unfortunately this interpretation is rarely accepted since y^* is an unobserved continuous surrogate of the ordinal outcome y.

Instead, we could determine the partial change in predicted probabilities:

$$\frac{\partial P(y = j|x, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\tau}_2, \widehat{\tau}_3)}{\partial x} = -\beta_1[f(\widehat{\tau}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x) - f(\widehat{\tau}_{j-1} - \widehat{\beta}_0 - \widehat{\beta}_1 x)].$$

This gives the slope of the curve relating x to $P(y = j|x)$:

$$Slope(j, x) = f(\widehat{\tau}_{j-1} - \widehat{\beta}_0 - \widehat{\beta}_1 x) - f(\widehat{\tau}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x).$$

You could make nice plots by keeping x fixed at some relevant value and varying j, or by keeping j fixed and varying x in some range of your choice. *Remark: such plots are rarely part of standard statistical packages. This is why you need to know the form of the density f (that is, the logistic density or the standard normal density) so you could create these plots on your own.*

Another interpretation is obtained by considering the odds of y being less than $j \in \{1, 2, 3, 4\}$ vs. greater than j:

$$\Omega_j(x) = \frac{P(y \leq j|x)}{P(y > j|x)} = \frac{F(\widehat{\tau}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x)}{1 - F(\widehat{\tau}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x)}.$$

For the ordered logit model, we have $F(x) = \text{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$ and we obtain (see also page 138 in Long's book):

$$\log \Omega_j(x) = \widehat{\tau}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x.$$

The interpretation of β_1 is as follows:

For a unit change in x, the odds $\Omega_j(x)$ are expected to change by a factor of $\exp(\widehat{\beta}_1)$.

Remark that the derivative says the factor should be $\exp(-\widehat{\beta}_1)$. However, since we do not specify the direction of the change, it is also fine to refer to $\exp(\widehat{\beta}_1)$.