

ORDERED LOGIT MODEL^{*}

Leonardo Grilli, Carla Rampichini

Dipartimento di Statistica, Informatica, Applicazioni “G. Parenti” – Università di Firenze

grilli@ds.unifi.it, rampichini@ds.unifi.it

The ordered logit model is a regression model for an ordinal response variable. The model is based on the cumulative probabilities of the response variable: in particular, the logit of each cumulative probability is assumed to be a linear function of the covariates with regression coefficients constant across response categories.

Questions relating to satisfaction with life assessment and expectations are usually ordinal in nature. For example, the answer to the question on how satisfied a person is with her quality of life can range from 1 to 10, with 1 being very dissatisfied and 10 being very satisfied (e.g. Schaafsma and Osoba, 1994; Anderson *et al.* 2009). It is tempting to analyse ordinal outcomes with the linear regression model, assuming equal distances between categories. However, this approach has several drawbacks which are well known in literature (see, for example, McKelvey and Zavoina, 1975; Winship and Mare, 1984; Lu, 1999). When the response variable of interest is ordinal, it is advisable to use a specific model such as the ordered logit model.

Let Y_i be an ordinal response variable with C categories for the i -th subject, alongside with a vector of covariates $\mathbf{x}_i$. A regression model establishes a relationship between the covariates and the set of probabilities of the categories $p_{ci} = \Pr(Y_i = y_c | \mathbf{x}_i)$, $c=1, \dots, C$. Usually, regression models for ordinal responses are not expressed in terms of probabilities of the categories, but they refer to convenient one-to-one transformations, such as the cumulative probabilities $g_{ci} = \Pr(Y_i \leq y_c | \mathbf{x}_i)$, $c=1, \dots, C$. Note that the last cumulative probability is necessarily equal to 1, so the model specifies only $C-1$ cumulative probabilities.

An ordered logit model for an ordinal response Y_i with C categories is defined by a set of $C-1$ equations where the cumulative probabilities $g_{ci} = \Pr(Y_i \leq y_c | \mathbf{x}_i)$ are related to a linear predictor $\beta' \mathbf{x}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots$ through the logit function:

$$\text{logit}(g_{ci}) = \log(g_{ci} / (1 - g_{ci})) = \alpha_c - \beta' \mathbf{x}_i, \quad c = 1, 2, \dots, C-1. \quad (1)$$

The parameters α_c , called *thresholds* or *cutpoints*, are in increasing order ($\alpha_1 < \alpha_2 < \dots < \alpha_{C-1}$). It is not possible to simultaneously estimate the overall intercept β_0 and all the $C-1$ thresholds: in fact, adding an arbitrary constant to the overall intercept β_0 can be counteracted by adding the same constant to each threshold α_c . This identification problem is usually solved by either omitting the overall constant from the linear predictor (i.e. $\beta_0 = 0$) or fixing the first threshold to zero (i.e. $\alpha_1 = 0$).

The vector of the slopes β is not indexed by the category index c , thus the effects of the covariates are constant across response categories. This feature is called the *parallel regression assumption*: indeed, plotting $\text{logit}(g_{ci})$ against a covariate yields $C-1$ parallel lines (or parallel curves in case of a non-linear

^{*} Draft of an entry of the Encyclopedia of Quality of Life Research, Michalos, Alex C. (Ed.). Springer. ISBN 978-94-007-0752-8.

specification, e.g. polynomial regression). In model (1) the minus before β implies that increasing a covariate with a positive slope is associated with a shift towards the right-end of the response scale, namely a rise of the probabilities of the higher categories. Some authors write the model with a plus before β : in that case the interpretation of the effects of the covariates is reversed.

From equation (1), the cumulative probability for category c is

$$g_{ci} = \exp(\alpha_c - \beta'x_i) / (1 + \exp(\alpha_c - \beta'x_i)) = 1 / (1 + \exp(-\alpha_c + \beta'x_i)) \quad (2)$$

The ordered logit model is also known as the *proportional odds model* because the parallel regression assumption implies the proportionality of the odds of not exceeding the c -th category $odds_{ci} = g_{ci} / (1 - g_{ci})$: in fact, the ratio of these odds for two units, say i and j , is $odds_{ci} / odds_{cj} = \exp[\beta'(x_j - x_i)]$, which does not depend on c and thus it is constant across response categories.

The ordered logit model is a member of the wider class of *cumulative ordinal models*, where the logit function is replaced by a general link function. The most common link functions are logit, probit and complementary log-log. These models are known in psychometrics as *graded response models* (Samejima, 1969) or *difference models* (Thissen and Steinberg, 1986). The last name indicates that the probabilities of the categories are obtained by difference: $p_{ci} = g_{ci} - g_{(c-1),i}$.

Early papers on regression models for ordinal data include McKelvey and Zavoina (1975), McCullagh (1980), and Winship and Mare (1984). The paper of Fullerton (2009) reviews ordered logistic regression models and their use in sociology. The textbook of Agresti (2010) gives a thorough treatment of ordinal data, while O'Connell (2006) provides applied researchers in the social sciences with accessible and comprehensive coverage of analyses for ordinal outcomes. Other valuable books fully devoted to ordinal outcomes are Johnson and Albert (1999) in a Bayesian perspective and Greene and Hensher (2010) in the setting of choice theory. Books on statistical modelling often have a chapter on ordinal regression models, for example Long (1997), Skrondal and Rabe-Hesketh (2004) and Hilbe (2009).

Representation as an underlying linear model with thresholds

An ordinal response Y_i with C categories can be represented as an underlying continuous response Y_i^* with a set of $C - 1$ thresholds α_c^* such that $Y_i = y_c$ if and only if $\alpha_{c-1}^* < Y_i^* \leq \alpha_c^*$. It follows that a cumulative model for an ordinal response, such as the ordered logit model (1), is equivalent to a system composed of a set of thresholds α_c^* and a linear regression model for an underlying continuous response:

$$Y_i^* = (\beta^*)' x_i + e_i^* \quad (3)$$

where e_i^* is an error with mean zero and standard deviation σ_{e^*} . The relationship $\Pr(Y_i \leq y_c) = \Pr(Y_i^* \leq \alpha_c^*)$ implies that the linear model (3) is equivalent to the cumulative model $l(g_{ci}) = \alpha_c - \beta'x_i$, where the link function $l(\cdot)$ is the inverse of the distribution function of the error e_i^* . The relationship between a parameter of the cumulative model θ and the corresponding parameter of the underlying model θ^* is $\theta = \theta^* \sigma_l / \sigma_{e^*}$, where σ_l is the standard deviation of the distribution associated to the link function (e.g. $\sigma_l = 1$ for probit and $\sigma_l = \pi/\sqrt{3} \cong 1.81$ for logit). Therefore, specifying the link function of the cumulative model amounts to specifying the distribution of the error of the underlying model and thus fixing its standard deviation to a conventional value: the probit corresponds to a standard normal error so the standard

deviation is fixed to 1, whereas the logit link corresponds to a standard logistic distribution so the standard deviation is fixed to $\pi/\sqrt{3} \cong 1.81$). Indeed, the measurement unit of the underlying model is undefined since $\Pr(Y_i^* \leq \alpha_c^*) = \Pr(kY_i^* \leq k\alpha_c^*)$ for any constant k , thus the standard deviation σ_{e^*} is not identifiable. This indeterminacy is solved in the cumulative model (1) since its parameters are measured on a conventional scale defined by the link (the standard deviation of the error does not appear as a parameter). The change of scale is the reason why the estimated regression coefficients from an ordered logit model are about 1.81 times the values from an ordered probit model.

The representation through an underlying linear model also makes clear that the estimated slopes from a cumulative model are approximately invariant to merging of the categories.

Relaxing the parallel regression assumption

The parallel regression assumption of the cumulative models may be too restrictive (for a test see Brant, 1990). Such an assumption can be relaxed by allowing the thresholds to depend on covariates or, alternatively, by allowing covariates to have category-specific slopes. These models are called *partial proportional odds* after Peterson and Harrell (1990). Another way to relax the parallel regression assumption is to let the variance of the disturbance e_i^* in the underlying linear model (3) to depend on covariates (McCullagh, 1980) or, alternatively, to use a scaled link such as the *scaled probit link* of Skrondal and Rabe-Hesketh (2004). A further approach is to introduce latent classes (Breen and Luijkx, 2010). Models violating the parallel regression assumption should be used with care since they raise identification and interpretation issues (Agresti, 2010).

Multilevel extension

Multilevel (random effects) ordered logit models are suitable for the analysis of correlated ordinal responses; see the reviews of Agresti and Natarajan (2001), Hedeker (2008) and Grilli and Rampichini (2011). Multilevel ordered logit or probit models may be useful in several kinds of applications in quality of life, for example: (i) analysis of a single response from individuals clustered into households, schools (e.g. Fielding *et al.*, 2003) or geographical regions (e.g. Rampichini and Schifini, 1998); (ii) joint analysis of a set of items of a survey questionnaire on individuals (e.g. Grilli and Rampichini, 2003); (iii) analysis of repeated responses to a given question in a longitudinal survey (e.g. Ribaud *et al.*, 1999).

References

1. Agresti, A (2010). *Analysis of Ordinal Categorical Data*, 2nd edition. New York: Wiley.
2. Agresti, A, Natarajan, R (2001). Modeling clustered ordered categorical data: A survey. *International Statistical Review*, 69: 345–371.
3. Anderson, R, Mikulić, B, Vermeylen, G, Lyly-Yrjanainen, M, Zigante, V (2009). *Second European Quality of Life Survey: Overview*. Luxembourg: Office for Official Publications of the European Communities.
4. Johnson, VE, Albert, JH (1999). *Ordinal Data Modeling*, New York: Springer.
5. Brant, R (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*. 46: 1171–1178.
6. Breen, R, Luijkx, R (2010). Mixture Models for Ordinal Data. *Sociological Methods & Research*. 39: 3–24.

7. Fielding, A, Yang, M, Goldstein, H (2003). Multilevel ordinal models for examination grades. *Statistical Modelling*, 3: 127–153.
8. Fullerton, AS (2009). A Conceptual Framework for Ordered Logistic Regression Models. *Sociological Methods & Research*, 38: 306–347.
9. Greene, WH, Hensher, DA (2010). *Modeling Ordered Choices: A Primer*. Cambridge UK: Cambridge University Press.
10. Grilli L, Rampichini C (2003). Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics*, 28: 31–44.
11. Grilli, L, Rampichini, C (2011). Multilevel models for ordinal data. In: Kenett R and Salini S (eds.) *Modern Analysis of Customer Satisfaction Surveys*. Wiley.
12. Hedeker, D (2008). Multilevel Models for Ordinal and Nominal Variables. In *Handbook of Multilevel Analysis* (ed. De Leeuw J and Meijer E), pp. 237–274. New York: Springer.
13. Hilbe, MH (2009). *Logistic Regression Models*, Chapman & Hall/CRC.
14. Johnson, VE, Albert, JH (1999). *Ordinal Data Modeling*, New York: Springer.
15. Long, S (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage.
16. Lu, M (1999). Determinants of Residential Satisfaction: Ordered Logit vs. Regression Models. *Growth and Change*, 30: 264–87.
17. McCullagh, P (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society Series B*, 42: 109–142.
18. McKelvey, RD, Zavoina, W (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4: 103–120.
19. O'Connell, A.A. (2006). *Logistic Regression Models for Ordinal Response Variables*. Sage.
20. Peterson, B, Harrell, FE (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39: 205–217.
21. Samejima, F (1969). Estimation of Latent Trait Ability Using A Response Pattern of Graded Scores. *Psychometric Monograph 17*, Psychometric Society, Bowling Green, OH.
22. Rampichini, C, Schifini, S (1998). A Hierarchical Ordinal Probit model for the analysis of life satisfaction in Italy, *Social Indicators Research*, 44: 5–39.
23. Ribaudo, HJ, Bacchi, M, Bernhard, J, Thompson, SG (1999). A Multilevel Analysis of Longitudinal Ordinal Data: Evaluation of the Level of Physical Performance of Women Receiving Adjuvant Therapy for Breast Cancer. *Journal of the Royal Statistical Society. Series A*, 162: 349–360.
24. Schaafsma, J, Osoba, D (1994). The Karnofsky Performance Status Scale Re-Examined: A Cross-Validation with the EORTC-C30, *Quality of Life Research*, 3: 413–424.
25. Skrondal, A, Rabe-Hesketh, S (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC Press, Boca Raton, FL.
26. Thissen, D, Steinberg, L (1986). A taxonomy of item response models. *Psychometrika*, 51: 567–577.
27. Winship, C, Mare, RD (1984). Regression models with ordinal variables. *American Sociological Review*, 49: 512–525.