In: Kenett R and Salini S (eds.) *Modern Analysis of Customer Surveys: with Applications using R*. Wiley, 2012.

19

Multilevel models for ordinal data

Leonardo Grilli and Carla Rampichini

Synopsis

This chapter is devoted to regression models for ordinal responses with special emphasis on random effects models for multilevel or clustered data. After a brief discussion on ordinal variables in the first section, the second section reviews the most common regression models for ordinal responses focusing on cumulative models, namely models based on cumulative probabilities. The third section deals with random effects cumulative models for multilevel data, discussing several issues peculiar to the random effects extension such as the distinction between marginal and conditional effects, the measures of unobserved cluster-level heterogeneity, the consequences of adding covariates, and the main types of predicted probabilities. The last part of the third section deals with estimation, inference and prediction, with a brief look on available software. The fourth section presents an application of random effects cumulative models to the analysis of student ratings on university courses.

Keywords: clustered data, correlated responses, cumulative model, mixed model, proportional odds model, random effects, unobserved heterogeneity.

19.1 Ordinal variables

Satisfaction is usually measured using graded scales, also called Likert scales, such as 'Very dissatisfied', 'Dissatisfied', 'Satisfied' and 'Very satisfied'. The resulting statistical variable Y is ordinal, namely it has ordered categories. Sometimes a score is associated with each label (e.g. '1: Very dissatisfied', '2: Dissatisfied', ...), but even in this case the variable Y is genuinely ordinal: it is not measured on an interval scale since the distances between the categories are unknown and the scoring system is just an arbitrary assumption. For example, the common choice of scoring the categories with the integers 1,2,3... amounts to assuming that the categories are evenly spaced (e.g. the difference between 'Very dissatisfied' and 'Dissatisfied' is the same as the difference between 'Dissatisfied' and 'Satisfied').

The statistical methods for ordinal variables avoid the arbitrariness of scoring systems and thus are generally to be preferred. Nonetheless, in the social sciences the use of scoring systems to convert categories into numbers is common practice since the statistical methods for quantitative variables are more powerful and easier to implement and interpret. The consequences of analyzing ordinal variables with methods for continuous variables have been investigated both analytically (Olsson 1979) and via simulations (Muthén and Kaplan 1985). In general, the bias depends on the number of categories (five is usually a minimum to get an acceptable bias) and the skewness of the distribution: indeed, the bias increases with the degree of skewness and may become large in the case of floor or ceiling effects, namely when the largest frequency corresponds to a category at the extremes of the scale. The bias may be reduced by using sophisticated scoring systems (Fielding 1997), but we do not pursue the matter further and later on we focus on the proper methods for ordinal variables.

An ordinal variable is a categorical variable supplemented with information on the ordering of the categories: indeed, the statistical methods for ordinal variables are designed to exploit such information. Formally, a categorical variable Y with categories y_c, $c = 1, \ldots, C$, has a multinomial distribution with probabilities $\pi_c = Pr(Y = y_c)$. The set of C probabilities $\pi_1, \pi_2, \ldots, \pi_C$ has one redundant probability due to the constraint $\pi_1 + \pi_2 + \ldots + \pi_C = 1$. When the categories are ordered, the cumulative probabilities are defined as $\gamma_c = Pr(Y \leq y_c) = \pi_1 + \pi_2 + \ldots + \pi_c$. Note that the non-redundant cumulative probabilities are $C - 1$ since the last one is $\gamma_C = 1$.

It is often useful to assume that an ordinal variable Y with C categories is generated by a latent continuous variable Y^* with a set of $C - 1$ thresholds α_c^* such that $Y = y_c$ if and only if $\alpha_{c-1}^* < Y^* \leq \alpha_c^*$. For example, if satisfaction is expressed using a 4-point scale (e.g. 'Very dissatisfied', 'Dissatisfied', 'Satisfied', 'Very satisfied'), we can postulate the existence of a latent satisfaction on a continuous scale which is categorized by 3 thresholds. Figure 19.1 represents the density of the underlying satisfaction Y^*, the thresholds α_c^* and the corresponding observed satisfaction Y.

The existence of an underlying continuous variable cannot be proved or disproved: it corresponds to a different approach useful for both interpretation and development of analytical tools, for example the polychoric correlations (e.g. Agresti 2010). The representation based on a latent continuous variable is conceptually appealing in settings such as customer satisfaction since it disentangles the process generating the observed satisfaction grade Y into two parts: the underlying satisfaction level Y^* and the measurement process corresponding to the thresholds α_c^*. In this perspective, the different ways of formulating the question and defining the rating scale affect the observed satisfaction grade through the thresholds. Similarly, if the rating scale adopted in the questionnaire is perceived in a different way by a subset of respondents, this affects the observed satisfaction grade through the thresholds. Indeed, the standard assumption that the set of thresholds is the same for all respondents is a measurement invariance assumption, which can be relaxed by allowing heterogenous thresholds (Johnson 2003).

In the next two sections we review the main regression models for an ordinal response: in section 19.2 we consider the standard, single-level models for independent observations, while in section 19.3 we consider the multilevel extension to deal with correlated observations. In single-level models (also called *marginal* models) the probabilities are conditioned on the covariates but not on the random effects, while in multilevel models (also called *conditional* models) the probabilities are conditioned both on the covariates and on

Figure 19.1 Density of underlying satisfaction, thresholds and observed satisfaction.

the random effects. In this chapter the probabilities and parameters for multilevel models have no superscript, while those for single-level models have the superscript $^\circ$ (for example, the vector of regression coefficients is $\boldsymbol{\beta}$ in a multilevel model and $\boldsymbol{\beta}^\circ$ in the corresponding single-level model). In all the models we will denote the statistical units by a subscript with double index ij, where $j = 1, 2, \ldots, J$ is the level 2 (cluster) index and $i = 1, 2, \ldots, n_j$ is the level 1 index (the double index is superfluous in single-level models, but we use it to keep the same notation across the chapter).

19.2 Standard models for ordinal data

We assume to observe an ordinal response Y_{ij} with C categories for level 1 unit i in cluster j, alongside with a vector of covariates $\mathbf{x}_{ij}$ (including the constant term). A regression model establishes a relationship between the covariates and the set of probabilities of the categories $\pi^\circ_{cij} = Pr(Y_{ij} = y_c|\mathbf{x}_{ij}), c = 1, \ldots, C$. Since one of the probabilities is redundant, any model must incorporate suitable restrictions to insure the identification of the parameters. Models for ordinal data also incorporate restrictions to reflect the ordering of the categories.

Models for ordinal data need not be expressed in terms of the set of category probabilities π°_{cij}: they may also refer to convenient one-to-one transformations, such as the set of cumulative probabilities $\gamma^\circ_{cij} = Pr(Y_{ij} \leq y_c|\mathbf{x}_{ij})$. Indeed, the most popular models for ordinal data are expressed in terms of these cumulative probabilities.

Early papers on regression models for ordinal data include McKelvey and Zavoina (1975), McCullagh (1980), and Winship and Mare (1984). The textbook of Agresti (2010) gives a thorough treatment of ordinal data, while O'Connell (2010) 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 Skrondal and Rabe-Hesketh (2004) and Hilbe (2009).

19.2.1 Cumulative models

A cumulative regression model for an ordinal response Y_{ij} with C categories is defined by a set of $C-1$ equations where the cumulative probabilities γ°_{cij} are related to the covariates $\mathbf{x}_{ij}$. We consider *cumulative generalised linear models* where the cumulative probabilities are related to the covariates through a linear predictor $\mathbf{x}'_{ij}\boldsymbol{\beta}^{\circ}$ and a monotone link function g:

$$g(\gamma^{\circ}_{cij}) = \alpha^{\circ}_{c} - \mathbf{x}'_{ij}\boldsymbol{\beta}^{\circ} \qquad c = 1, 2, \ldots, C-1. \tag{19.1}$$

The parameters α°_{c}, called *thresholds* or *cutpoints*, are in increasing order, $\alpha^{\circ}_{1} < \alpha^{\circ}_{2} < \ldots < \alpha^{\circ}_{C-1}$. The vector of regression coefficients $\boldsymbol{\beta}^{\circ}$ (including the intercept β°_{0}) does not have the category index c, thus the effects of the covariates are constant across response categories, a feature called the *parallel regression assumption*: indeed, plotting $g(\gamma^{\circ}_{cij})$ against a covariate yields $C-1$ parallel lines (or parallel curves if the covariate has non-linear terms). Cumulative 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: $\pi^{\circ}_{cij} = \gamma^{\circ}_{cij} - \gamma^{\circ}_{c-1,ij} = g^{-1}(\alpha^{\circ}_{c} - \mathbf{x}'_{ij}\boldsymbol{\beta}^{\circ}) - g^{-1}(\alpha^{\circ}_{c-1} - \mathbf{x}'_{ij}\boldsymbol{\beta}^{\circ})$.

The minus sign before the linear predictor in model (19.1) 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 the linear predictor: in that case the interpretation of the effects of the covariates is reversed.

In model (19.1) we cannot simultaneously estimate the constant of the linear predictor and all the $C-1$ thresholds: in fact, adding an arbitrary constant to the linear predictor can be counteracted by adding the same constant to each threshold. This identification problem is usually solved by either omitting the constant from the linear predictor ($\beta^{\circ}_{0} = 0$) or fixing the first threshold to zero ($\alpha^{\circ}_{1} = 0$).

Typical choices of the link function g are *logit*, *probit* and *complementary log-log*. The logit link is widely used (except in the social sciences) mainly because of the connection with odds ratios. In the following we will focus on the logit cumulative model, also known as the *proportional odds model*:

$$logit(\gamma^{\circ}_{cij}) = \alpha^{\circ}_{c} - \mathbf{x}'_{ij}\boldsymbol{\beta}^{\circ} \qquad c = 1, 2, \ldots, C-1, \tag{19.2}$$

where the logit on the left hand side is a *cumulative logit*, namely the logarithm of the odds of not exceeding the c-th category:

$$logit(\gamma^{\circ}_{cij}) = log\frac{\gamma^{\circ}_{cij}}{1-\gamma^{\circ}_{cij}} = log\frac{Pr(Y_{ij} \leq y_c)}{Pr(Y_{ij} > y_c)}\,. \tag{19.3}$$

In model (19.2) the parallel regression assumption implies the proportional odds property: in fact, the ratio of the odds of not exceeding the c-th category for units ij and $i'j'$ is $exp(-(\mathbf{x}'_{ij} - \mathbf{x}'_{i'j'})\boldsymbol{\beta}^{\circ})$, which does not depend on c and thus is constant across response categories.

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 residual in the underlying linear model (see subsection 19.3.1) to depend on covariates (Cox 1995) 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).

19.2.2 Other models

Even if the rest of this chapter will be devoted to the multilevel extension of cumulative models, we briefly mention some non-cumulative models that may be preferable in some contexts and can be extended to handle multilevel data as well.

A wide class of models is obtained by specifying a multinomial logit model for the probabilities of the categories $\pi^{\circ}_{cij} = Pr(Y_{ij} = y_c|\mathbf{x}_{ij})$ with additional parameter constraints reflecting the ordering of the categories (Skrondal and Rabe-Hesketh 2004). An example is the *adjacent category logit model* (Agresti 2010), where the linear predictor is equated to the logarithm of the odds between adjacent categories $\pi^{\circ}_{cij}/\pi^{\circ}_{c-1,ij}$. In fact, such a model can be written as a multinomial logit model with linear predictor $c\mathbf{x}'_{ij}\boldsymbol{\beta}^{\circ}$, which can be seen as either a model with category-specific slopes $c\boldsymbol{\beta}^{\circ}$ and category-invariant covariates $\mathbf{x}_{ij}$ or a model with category-invariant slopes $\boldsymbol{\beta}^{\circ}$ and category-specific covariates $\mathbf{x}^{c}_{ij} = c\mathbf{x}_{ij}$. The last formulation is known in Item Response Theory as the *partial credit model* (Masters 1982), which is a generalization of the Rasch model to ordinal items.

A valuable alternative to traditional models for ordinal responses is represented by the CUB models outlined in Chapter 13. While traditional models considered in this chapter are based on a multinomial distribution, CUB models are based on a mixture between a shifted binomial distribution (to be interpreted as *feeling*) and a discrete uniform distribution (to be interpreted as *uncertainty*).

19.3 Multilevel models for ordinal data

Let us now consider the multilevel extension of regression models for ordinal responses. These models are outlined in most books on multilevel analysis. In addition, we recommend the reviews of Agresti and Natarajan (2001) and Hedeker (2008), and chapter 10 of Agresti (2010). As the field is vast, we focus on the most popular configuration in applications, namely cumulative models (outlined in subsection 19.2.1) with a random intercept in a two-level hierarchy:

$$g(\gamma_{cij}) = \alpha_c - (\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j) \qquad c = 1, 2, \ldots, C-1, \tag{19.4}$$

where $j = 1, 2, \ldots, J$ is the level 2 (cluster) index and $i = 1, 2, \ldots, n_j$ is the level 1 index, and γ_{cij} is the cumulative probability up to the c-th category for unit i in cluster j. The term u_j is a random effect representing unobserved factors at the cluster level: since it is

shared by all the units of the cluster, it induces within-cluster correlated responses. If the overall intercept β_0 is unconstrained, we can view u_j as a random shift of the intercept so that the intercept of cluster j is $\beta_0 + u_j$; otherwise, if the overall intercept is fixed to zero, we can view u_j as a random shift of the thresholds so that the set of thresholds of cluster j is $\alpha_c - u_j, c = 1, 2, \ldots, C-1$.

The standard assumption on the random effects u_j is that, conditionally on the covariates, they are independent and identically distributed with zero mean and a common cluster variance σ_u^2 to be estimated. On the contrary, the assumption of common cluster variance can be easily relaxed (Hedeker (2008), section 6.7), as well as the conventional normality assumption (Agresti and Natarajan (2001), section 4.2). In order to get unbiased estimates, the key part of the standard assumption is the *exogeneity*, namely the mean of the random effects does not depend on the covariates: $E(u_j|\{\mathbf{x}_{ij} : i = 1, 2, \ldots, n_j\}) = 0$. A multilevel model like the one in equation (19.4) may be useful in several kinds of applications in customer satisfaction, for example: (*i*) analysis of a single response from customers clustered in units offering a product or service (firms, schools, hospitals ...) or clustered in geographical regions; (*ii*) analysis of repeated responses to a given question in a longitudinal survey on a panel of customers; (*iii*) joint analysis of a set of items of a survey questionnaire on customers. Note that customers are level 1 units in example (*i*) and level 2 units (clusters) in examples (*ii*) and (*iii*).

The sample size required for reliably fitting a multilevel model for ordinal data depends on several factors, including the complexity of the model, the value of the cluster variance and the estimation method. Moreover, the requirement is higher for the variances of the random effects than for the regression coefficients. Some guidelines are provided by recent simulation studies on the closely related multilevel logit models for binary responses: Austin (2010) considers a random intercept logit model, whereas Moineddin *et al.* (2007) focus on a logit model where both the intercept and the slope randomly vary across clusters. In the random intercept case, the estimates are reasonably good with most estimation methods even with 10 to 15 clusters as long as the average cluster size is at least 10. If the clusters are smaller, more clusters are needed. In the random slope case, the requirement is considerably higher, say 30 clusters of size 30.

19.3.1 Representation as an underlying linear model with thresholds

As noted in section 19.1, an ordinal response Y_{ij} with C categories can be represented as an underlying continuous response Y_{ij}^* with a set of $C-1$ thresholds α_c^* such that $Y_{ij} = y_c$ if and only if $\alpha_{c-1}^* < Y_{ij}^* \leq \alpha_c^*$. It follows that a cumulative generalised linear model for an ordinal response is equivalent to a system composed of a set of thresholds α_c^* and a linear regression model for an underlying continuous response:

$$Y_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta}^* + u_j^* + e_{ij}^*\,, \tag{19.5}$$

where e_{ij}^* is a level 1 error with standard deviation σ_{e^*} and u_j^* is a level 2 error with standard deviation σ_{u^*}. In fact,

$$Pr(Y_{ij} \leq y_c) = Pr(Y_{ij}^* \leq \alpha_c^*) = Pr(e_{ij}^* \leq \alpha_c^* - \mathbf{x}_{ij}'\boldsymbol{\beta}^* - u_j^*) = g^{-1}(\alpha_c - \mathbf{x}_{ij}'\boldsymbol{\beta} - u_j).$$

Therefore, the underlying linear model (19.5) with thresholds α_c^* and level 1 error e_{ij}^* having distribution function g^{-1} is equivalent to the cumulative model (19.4) with link function

g. The relationship between a parameter of the cumulative model θ and the corresponding parameter of the underlying model θ^* is $\theta = \theta^* \sigma_g / \sigma_{e^*}$, where σ_g is the standard deviation of the distribution associated to the link function (e.g. $\sigma_g = 1$ for probit and $\sigma_g = \pi/\sqrt{3} \simeq 1.81$ for logit).

When we specify the link function for the cumulative model, we implicitly specify the distribution function of the level 1 error and, consequently, we fix the standard deviation of the level 1 error to a conventional value: the probit link corresponds to a standard normal error so the standard deviation is fixed to 1, whereas the logit link corresponds to a standard logistic error so the standard deviation is fixed to $\pi/\sqrt{3} \simeq 1.81$. Indeed, the measurement unit of the underlying model is undefined since $Pr(Y^*_{ij} \leq \alpha^*_c) = Pr(kY^*_{ij} \leq k\alpha^*_c)$ for any constant k, thus the standard deviation σ_{e^*} is not identifiable. This indeterminacy is solved in the cumulative model (19.4) since its parameters are measured on a conventional scale defined by the link (the level 1 standard deviation does not appear as parameter). The change of scale is the reason why replacing probit with logit causes an expansion of the estimated slopes of about 1.81. The model specification requires some care in case of level 1 heteroscedasticity, for example when σ_{e^*} changes across strata (Grilli and Rampichini 2002).

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

19.3.2 Marginal versus conditional effects

The slopes $\boldsymbol{\beta}$ of the random intercept cumulative model (19.4) represent *conditional* or *cluster-specific* effects: they summarize the relationship between the covariates $\mathbf{x}_{ij}$ and the conditional cumulative probabilities $\gamma_{cij} = Pr(Y_{ij} \leq y_c | \mathbf{x}_{ij}, u_j)$, which are conditional on the random effect and thus refer to a specific cluster of the population. On the other hand, the slopes $\boldsymbol{\beta}^\circ$ of the standard cumulative model (19.1) represent *marginal* or *population-averaged* effects: they summarize the relationship between the covariates $\mathbf{x}_{ij}$ and the marginal cumulative probabilities $\gamma^\circ_{cij} = Pr(Y_{ij} \leq y_c | \mathbf{x}_{ij})$, which are marginal with respect to the random effect and thus refer to the whole population.

Marginal effects are smaller in absolute value than conditional effects, namely $|\beta^\circ_m| \leq |\beta_m|$ for every covariate. Such attenuation can be shown using the representation with the underlying linear model. In fact, in subsection 19.3.1 we showed that the m-th slope of the random intercept cumulative model (19.4) is $\beta_m = \beta^*_m \sigma_g / \sigma_{e^*}$; on the other hand, if the random effect u^*_j is omitted, the underlying linear model (19.5) has a composite level 1 error $d^*_{ij} = u^*_j + e^*_{ij}$ with standard deviation $\sigma_{d^*} = \sqrt{\sigma^2_{u^*} + \sigma^2_{e^*}}$, thus the corresponding slope of the single-level cumulative model (19.1) is $\beta^\circ_m = \beta^*_m \sigma_g / \sigma_{d^*}$. Since $\sigma_{d^*} \geq \sigma_{e^*}$ it follows that $|\beta^\circ_m| \leq |\beta_m|$ for every covariate. Clearly, the attenuation is stronger the larger is the level 2 variance compared to the level 1 variance, namely the higher is the unobserved heterogeneity due to the clustering of the units. Under the standard assumption of a normal random effect, the analytical development outlined above is exact in the probit case (since a random intercept probit model implies a marginal probit model) and approximate in general (for example, a random intercept logit model does not imply a marginal logit model).

Marginal and conditional slopes are population parameters: regardless of the estimation methods, a model without random effects has marginal slopes, while a model with random effects has conditional slopes. In most applications, conditional slopes are of interest as they refer to the cluster-specific effects, which are more informative about causal processes.

Finally, note that if the responses are correlated within the clusters, random effects models yield correct standard errors, while marginal models, namely without random effects, yield wrong standard errors (usually underestimated). Thus, if one is interested in marginal effects in presence of correlated data, two alternatives are possible: (1) fit a random effects model and then recover the marginal effects, or (2) fit a marginal model using a correction for the standard errors, such as the GEE method or a robust estimator of the standard errors (Agresti and Natarajan 2001).

19.3.3 Summarizing the cluster-level unobserved heterogeneity

In a linear random intercept model like (19.5) the level of unobserved heterogeneity due to the clustering of the units is summarized by the Intraclass Correlation Coefficient (ICC) $\rho = \sigma^2_{u^*}/(\sigma^2_{u^*} + \sigma^2_{e^*})$. In a linear model the ICC is both the proportion of the between-cluster variance with respect to the total variance and the correlation between the responses of two units of the same cluster, namely $\rho = Cor(Y^*_{ij}, Y^*_{i'j}|\mathbf{x}_{ij}, \mathbf{x}_{i'j})$. Such a correlation does not depend on the covariates (it is homogeneous), so the ICC is an exhaustive indicator of the degree of correlation. Unfortunately, this property does not hold in models for categorical responses such as the random intercept cumulative model (19.4) since $Cor(Y_{ij}, Y_{i'j}|\mathbf{x}_{ij}, \mathbf{x}_{i'j})$ actually depends on the covariates. An appealing solution is to summarize the degree of within-cluster correlation using the ICC for the underlying linear model, which can be easily computed using the cluster variance σ^2_u of the cumulative model: in fact, from the relationship $\sigma_u = \sigma_{u^*}\sigma_g/\sigma_{e^*}$ of subsection 19.3.1, it follows that $\rho = \sigma^2_u/(\sigma^2_u + \sigma^2_g)$, where σ^2_g is the variance of the distribution associated to the link function. For example, $\rho = \sigma^2_u/(\sigma^2_u + 1)$ for probit and $\rho = \sigma^2_u/(\sigma^2_u + \pi^2/3)$ for logit. However, the ICC for the underlying linear model is misleading if one attempts to compare it with the values usually obtained in linear models for observed continuous responses: in fact, the ICC for the underlying linear model is much lower and it often gives the impression of a negligible within-cluster correlation. For example, a value $\rho = 0.01$ is negligible for an observed response but not for an underlying response.

A simple and effective way of summarizing the within-cluster correlation in models for categorical responses is to compute the probabilities under several scenarios defined by fixing the random effect u_j to a set of percentiles of its estimated distribution. Denoting with $u_{[p]}$ the percentile p, if the random intercept cumulative model (19.4) has a normally distributed u_j with estimated standard deviation $\widehat{\sigma}_u$, then three scenarios could be defined by fixing the random effect to the estimated percentiles $\widehat{u}_{[2.5]} = -1.96\widehat{\sigma}_u$, $\widehat{u}_{[50]} = 0$ and $\widehat{u}_{[97.5]} = +1.96\widehat{\sigma}_u$. Once the covariates have been fixed to a set of values $\mathbf{x}^0$, the cumulative probability up to category c in the scenario corresponding to percentile p is defined as $Pr(Y \leq y_c|\mathbf{x}^0, \widehat{u}_{[p]})$ and it is computed by replacing the model parameters with their estimates.

19.3.4 Consequences of adding a covariate

The representation of a cumulative model for ordinal responses as an underlying linear model with thresholds (subsection 19.3.1) shows that the estimable parameters are scaled by the underlying level 1 standard deviation σ_{e^*}, for example $\beta_m = \beta^*_m\sigma_g/\sigma_{e^*}$. If it were possible to observe Y^*_{ij} and fit the underlying linear model (19.5), the addition of a covariate would

reduce σ_{e^*}. However, in a cumulative model for the observable ordinal response Y_{ij} the level 1 standard deviation σ_{e^*} cannot change, so the effect on σ_{e^*} is dumped on the other parameters. This phenomenon can be easily seen in the hypothetical case of a new covariate w_{ij} with the following features: (*i*) w_{ij} is independent of the other covariates, so its inclusion does not alter the slopes $\boldsymbol{\beta}^*$ of the other covariates; (*ii*) w_{ij} has no between-cluster variation, so its inclusion does not alter the level 2 standard deviation σ_{u^*}; (*iii*) w_{ij} has some within-cluster variation, so its inclusion reduces the level 1 standard deviation to $k\sigma_{e^*}$ with $k < 1$. It follows that the addition of the new covariate w_{ij} in the cumulative model (19.4) inflates all the parameters by $1/k$, for example $\beta_{m,new} = \frac{\beta^*_m \sigma_g}{k\sigma_{e^*}} = \frac{1}{k}\beta_{m,old}$. Note that also the cluster variance increases, a phenomenon that may appear surprising since the added covariate has no between-cluster variation.

The simple pattern outlined for the hypothetical covariate w_{ij} does not hold in general, but it is clear that the change of scale induced by the new covariate hinders a direct comparison of the parameters before and after its inclusion. This issue (which concerns also the models for binary responses) is considered by Winship and Mare (1984) in the case of single-level models and by Fielding (2004) and Bauer (2009) in the case of multilevel models.

19.3.5 Predicted probabilities

In multilevel models for categorical responses there are three types of predicted probability (Skrondal and Rabe-Hesketh 2009): (*i*) *conditional probability* (a unit in a hypothetical cluster); (*ii*) *population-averaged probability* (a unit in a new, randomly sampled cluster); (*iii*) *cluster-averaged probability* (a unit in a specific cluster of the sample). All these types of predicted probability require to replace the parameters with their estimates and the covariates with arbitrary values. The three types of probability differ in the way the random effect u_j is handled: in the conditional probability, the random effect u_j is fixed to an arbitrary value (usually chosen as a percentile of its estimated distribution in the population, see subsection 19.3.3); in the other instances, the random effect u_j is averaged out using its estimated distribution in the whole population (population-averaged type) or using its estimated distribution for the j-th cluster of the sample (cluster-average type). Skrondal and Rabe-Hesketh (2009) give guidelines on computation and interpretation.

Predicted probabilities are essential for an effective and intelligible report of the model results. In the random intercept cumulative model (19.4) the effects of the covariates are summarized by the vector of the slopes $\boldsymbol{\beta}$. Unfortunately, the interpretation of $\boldsymbol{\beta}$ is not straightforward as it refers to a transformation g of the cumulative probabilities. A consequence is that, like in any model for categorical responses, the change in a certain probability due to a unit increase in a covariate depends on the value of such a probability (the closer the probability to 0 or 1, the smaller the change). It is therefore important to express the effects of the covariates in terms of changes in the predicted probabilities with reference to some relevant scenarios. A popular strategy in a model with M covariates is to compute $M + 1$ sets of predicted probabilities of the categories $\widehat{\pi}_1^{(m)}, \widehat{\pi}_2^{(m)}, \ldots, \widehat{\pi}_C^{(m)}$ ($m = 0, 1, \ldots, M$), where the set $m = 0$ refers to a hypothetical baseline subject and the other sets consider a unit increase in the m-th covariate.

19.3.6 Cluster-level covariates and contextual effects

As in any multilevel model, the covariates $\mathbf{x}_{ij}$ of the random intercept cumulative model (19.4) can include cluster-level covariates and cross-level interactions. The multilevel analysis with an ordinal response follows the basic principles explained in the textbooks with reference to linear models for continuous responses (e.g. Raudenbush and Bryk 2002), even if there are some complications. For example, we noted in subsection 19.3.1 that the level 1 and level 2 variances cannot be estimated separately since only their ratio is identified. As a consequence, the level 2 variance of the cumulative model may increase after the inclusion of a covariate with no cluster-level variation (see subsection 19.3.4).

Another complication with categorical responses is related to the assessment of a contextual effect, which is a key quantity in education and sociology (Raudenbush and Bryk 2002). In a linear model for a continuous response, the *contextual effect* of a covariate z_{ij} is the coefficient δ of its cluster mean $\overline{z}_j$ when both z_{ij} and $\overline{z}_j$ enter as covariates. Thus δ is the change in the expectation of the response following a unit increase in the cluster mean $\overline{z}_j$ while keeping constant the individual-level covariate z_{ij}. In a linear model, the change in the expectation of the response does not depend on the values of $\overline{z}_j$ and z_{ij}, so the contextual effect is a unique value denoted by the parameter δ. However, in a model for categorical responses, δ is just the contextual effect on the scale of the linear predictor: in order to assess the contextual effect on the probabilities, it is necessary to compute the predicted probabilities under several scenarios and make plots. This approach is illustrated in Skrondal and Rabe-Hesketh (2009).

19.3.7 Estimation of model parameters

The estimation of the parameters of the random intercept cumulative model (19.4) is usually based on maximum-likelihood, yielding unbiased estimates under the missing at random assumption (MAR, Rubin 1976); that is, the missingness mechanism may depend on both model covariates and observed responses. Under mild regularity conditions, ML estimators have good asymptotic properties, e.g. consistency, normality and efficiency. In this framework the asymptotic theory requires increasing the number of clusters (increasing the cluster sizes is not enough), so the number of clusters J is the key quantity for asymptotics.

Let $\mathbf{Y}_j$ be the vector of the n_j ordinal responses of the j-th cluster and let $\mathbf{X}_j$ be the covariate matrix of cluster j with rows $\mathbf{x}'_{ij}$. Moreover, let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_u)'$ be the vector of the model parameters, where $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_{C-1})$. The likelihood of $\mathbf{Y}_j$ conditional on u_j is equal to the product of the conditional probabilities of the responses

$$L_j(\mathbf{Y}_j \mid u_j, \mathbf{X}_j; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n_j} \prod_{c=1}^{C} \pi_{cij}^{d_{cij}} = \prod_{i=1}^{n_j} \prod_{c=1}^{C} (\gamma_{cij} - \gamma_{c-1,ij})^{d_{cij}}$$

where d_{cij} is the indicator of $\{Y_{ij} = y_c\}$, whereas $\gamma_{0,ij} = 0$ and $\gamma_{C,ij} = 1$.

The likelihood of $\mathbf{Y}_j$ is obtained integrating out the (unobservable) random effect u_j, i.e.

$$L_j(\mathbf{Y}_j \mid \mathbf{X}_j; \boldsymbol{\theta}) = \int_{u_j} L_j(\mathbf{Y}_j \mid u_j, \mathbf{X}_j; \boldsymbol{\alpha}, \boldsymbol{\beta}) f(u_j; \sigma_u) du_j \tag{19.6}$$

where $f(u_j; \sigma_u)$ is the density of u_j, usually assumed to be a normal density with zero mean and standard deviation σ_u.

Given independence across clusters, the log-likelihood for the J clusters is

$$\log L = \sum_{j=1}^{J} \log L_j(\mathbf{Y}_j \mid \mathbf{X}_j; \boldsymbol{\theta})$$

which is maximized to obtain ML estimates of the model parameters.

In general, the integral in the likelihood (19.6) is not in closed form, thus some type of approximation is needed. Various approaches have been proposed in the literature, including Gaussian quadrature, Laplace approximation and Monte Carlo integration. Reviews are given by Skrondal and Rabe-Hesketh (2004) and Hedeker (2008). Each technique has advantages and drawbacks in terms of precision and computational burden. Different techniques usually yield slightly different parameter estimates, especially for the variance-covariance parameters of the random effects.

The most widely used technique for numerical integration is Gaussian quadrature, which can be ordinary, adaptive or spherical. The precision of the estimates and the computational burden depend heavily on the number of quadrature points: more quadrature points imply higher accuracy and longer computational time. The computational burden, which may be prohibitive in some cases, also depends on factors such as the number of observations and the number of random effects (dimensionality of the integral). Simulations show that the adaptive version of Gaussian quadrature performs well in a wide variety of situations as long as the dimensionality of the integral does not exceed 5 or 6 (Rabe-Hesketh *et al.* 2005).

The 6th order Laplace approximation by Raudenbush *et al.* (2000) appears to be very efficient and sufficiently accurate in many situations (Joe 2008). Muthén's limited information approach is an excellent alternative for models with multivariate normal random effects when the cluster sizes are nearly constant and there are few missing data (Muthén and Satorra 1995).

Quasi-likelihood methods, such as MQL and PQL, are based on first- or second-order Taylor approximation of the likelihood. They are computationally efficient but, in some situations, they underestimate the cluster variance and thus they yield attenuated slopes (Mealli and Rampichini 1999; Rodriguez and Goldman 1995). MQL is faster but more biased than PQL. Both methods have the drawback of preventing likelihood-based inference since the likelihood function is not evaluated.

Standard likelihood-based methods require to specify a parametric continuous distribution for the random effects, typically the Normal distribution. Alternatively, it is possible to specify an arbitrary discrete distribution and estimate both the locations and the masses (Aitkin 1999). The model with discrete random effects is also called 'finite mixture' or 'latent class'. The latent class is a set of clusters, which is latent because the membership of clusters to classes is unobservable. Each class is characterized by a (prior) probability and a location for the random effect. For a fixed number of mass points, the estimation is straightforward since the likelihood is a finite mixture and no integration is involved. However, choosing the number of mass points is a difficult task, since the comparison of models with different numbers of mass points cannot be done with standard likelihood-based tests. A practical solution is to compare the models using fit indexes such as AIC and BIC (with many variations) or the Gateaux derivative method (Rabe-Hesketh *et al.* 2003). The resulting estimator is known as Non-Parametric Maximum Likelihood (NPML, Lindsay *et al.* 1991).

In the Bayesian approach to multilevel models, both fixed and random effects are considered to be random variables with a given prior distribution and inference is based on their joint posterior distribution. This approach is more demanding than ML since it requires to specify the prior distribution of the model parameters and to use computationally intensive MCMC algorithms. The effort may be worthwhile in highly complex models since the Bayesian approach is better than ML in assessing the uncertainty of the estimates, a feature that may have considerable consequences on the coverage of confidence intervals (Browne and Draper 2006). Moreover, Bayesian methods do not rely on asymptotics, thus outperforming ML in small samples (Austin 2010).

19.3.8 Inference on model parameters

Standard large-sample inference procedures are applicable when the model is estimated via ML methods. Hypothesis testing for the fixed-effects parameters (i.e., $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) can be conducted in the usual way, using the Wald test or the likelihood ratio test (LRT).

Inference about the cluster variance requires some care. In fact, unless the number of clusters is very large, the Wald test should not be used since the sampling distribution of the estimator of the cluster variance is highly asymmetric. The LRT is preferable. However, the null hypothesis of main interest, namely $\sigma_u^2 = 0$, is on the boundary of the parameter space and thus standard asymptotic results do not hold for the test statistics, including LRT. Indeed, the asymptotic distribution of the LRT statistic for $\sigma_u^2 = 0$ is not a chi-square with 1 d.f., but rather it is a 50:50 mixture of a mass point at 0 and a chi-square with 1 d.f. (Berkhof and Snijders 2001; Stram and Lee 1994). A practical solution is to perform the usual LRT and then halve the p-value (otherwise the test is conservative, i.e. the actual probability of type I error is lower than the nominal level). Alternatively, Verbeke and Molenberghs (2003) derived general one-sided score tests for variance components in models with several random effects.

19.3.9 Prediction of random effects

In many cases, it is useful to assign values to the random effects. Predicted random effects can be used for inference regarding clusters, for example to assess effectiveness of schools, universities, hospitals or firms (Grilli and Rampichini 2009). Moreover, predicted random effects are useful quantities in model diagnostics, e.g. to check for violations of the normality assumption for random effects or to find outliers (Snijders and Berkhof 2008).

The u_j are usually predicted using Empirical Bayes (EB) methods (Skrondal and Rabe-Hesketh 2009). In this setting the population distribution of the random effects is called *prior*, whereas the distribution of the random effects conditional on the data of a given cluster is called *posterior*. The EB prediction is the mean of the posterior distribution with parameter estimates plugged in, combining data information (likelihood) with population information (prior),

$$\hat{u}_j^{EB} = E(u_j \mid \mathbf{Y}_j, \mathbf{X}_j; \widehat{\boldsymbol{\theta}}) = \int u_j h(u_j \mid \mathbf{Y}_j, \mathbf{X}_j; \widehat{\boldsymbol{\theta}}) du_j, \tag{19.7}$$

where $h(\cdot)$ is the empirical posterior distribution of u_j. The mean of the posterior distribution is a value between 0 (the mean of the prior) and the mode of the likelihood of the j-th cluster: the prediction that would be obtained using only the cluster-specific likelihood is

thus shrunken, with a stronger shrinkage for small clusters. The EB predictor is conditionally biased towards zero and unconditionally unbiased (Skrondal and Rabe-Hesketh 2009).

In the multilevel ordinal model (19.4) the EB predictions (19.7) do not have closed form, thus numerical or simulation-based integration methods must be used.

An alternative way to assign values to the random effects uses the posterior mode of the random effects. EB modal predictions do not require numerical integration.

Note that EB predictions are a by-product of the MLE algorithms relying on adaptive quadrature. For example, the `gllamm` procedure of Stata yields posterior means (Rabe-Hesketh *et al.* 2005), while the routines implemented in `R` yield posterior modes (Pinheiro and Bates 1995).

There are two kinds of standard errors of the EB predictions (19.7), depending on their use. *Comparative standard errors* are used for inference regarding the *true* values of u_j for specific clusters (e.g. for making comparisons between clusters, see Snijders and Bosker 1999); on the other hand, the *diagnostic standard errors* are useful for model diagnostics (e.g. for finding outliers, see Snijders and Berkhof 2008).

The comparative standard error is the square root of the posterior variance,

$$var(u_j \mid \mathbf{Y}_j, \mathbf{X}_j; \widehat{\boldsymbol{\theta}}) = \int (u_j - \hat{u}_j^{EB})^2 h(u_j \mid \mathbf{Y}_j, \mathbf{X}_j; \widehat{\boldsymbol{\theta}}) du_j, \tag{19.8}$$

which has no closed form and thus the integral in (19.8) must be approximated.

For model diagnostics it is useful to consider the marginal sampling variance of the EB predictor, i.e. the variance of the prediction under repeated sampling of the responses from their marginal distribution, keeping the covariates fixed and plugging in parameter estimates. There is no closed form expression of the marginal sampling variance in the ordinal multilevel model. Skrondal and Rabe-Hesketh (2009) suggest the following approximation

$$var(\hat{u}_j^{EB} \mid \mathbf{X}_j; \widehat{\boldsymbol{\theta}}) \approx \hat{\sigma}_u^2 - var(u_j \mid \mathbf{Y}_j, \mathbf{X}_j; \widehat{\boldsymbol{\theta}}).$$

Both the posterior standard deviation (*comparative standard error*) and the sampling standard deviation (*diagnostic standard error*) of the EB prediction are lower than the estimated prior standard deviation $\widehat{\sigma}_u$. The posterior standard deviation tends to decrease as the cluster size n_j becomes larger, reflecting the increasing accuracy in the prediction of u_j. On the contrary, the sampling standard deviation increases with n_j because the EB prediction is less shrunken.

On the basis of a simulation study, Skrondal and Rabe-Hesketh (2009) recommend using the posterior standard deviation as comparative standard error, while they find that the sampling distribution of the empirical Bayes predictions is often too discrete and non-normal for the diagnostic standard error to be used in the usual way for identifying outliers.

19.3.10 Software

Multilevel models for ordinal data can be fitted with ML or Bayesian methods using procedures in general purpose statistical packages (e.g. R, SAS and Stata), specialized software for multilevel analysis (e.g. MLwiN and HLM) or specialized software for latent variable models (e.g. Mplus and Latent GOLD). The programs are different in many respects. In particular, it is important to bear in mind that the parameter estimates may change with the method used to numerically evaluate the likelihood.

Multilevel modelling software reviews are available at the web site of the Centre for Multilevel Modelling in Bristol.

In the following we list the programs with special emphasis on those implementing ML via numerical integration, giving references for more details. The list is not complete and rely mainly on our personal experience.

Software for ML estimation

Multilevel ordinal models can be fitted with ML by several programs. Most programs perform ML estimation via numerical integration, often using some form of quadrature.

The `ordinal` package of `R` (Christensen 2010) fits cumulative link mixed models for ordinal data, though it is limited to random intercept models. The package includes the proportional odds model but it also allows for general regression structures for the location and the scale of the latent distribution (additive and multiplicative structures, structured thresholds and flexible link functions). Furthermore, several estimation procedures and auxiliary functions are implemented.

PROC NLMIXED of SAS (SAS 2009) is a general routine for non-linear mixed models. Multilevel ordinal models can be estimated by writing down the model likelihood using SAS statements. The procedure offers a wide choice of integral approximations and optimization techniques.

The `gllamm` command (Rabe-Hesketh *et al.* 2008) of Stata provides tools for analyzing multilevel ordinal data. This procedure fits models of the GLLAMM class (Generalized Linear Latent And Mixed Models) by ML with several kinds of quadrature. Moreover, it allows to relax the parallel regression assumption (see section 19.2.1) by specifying a model for the thresholds or by using a scaled probit link.

The freeware program MIXOR provides ML estimates for mixed effects ordinal regression models (Hedeker and Gibbons 1996). The commercial version is implemented in the program SUPERMIX (Hedeker and Gibbons 2008).

Another freeware software for mixed effects ordinal models is aML, which is a general software for multilevel, multiprocess models (Lillard and Panis 2003).

ML estimates of multilevel ordinal models via numerical integration are also provided by programs for latent variable models, such as Latent GOLD (Vermunt and Magidson 2005) and Mplus (Muthén and Muthén 2010).

Peculiar estimation techniques are available in the specialized multilevel software HLM (Raudenbush *et al.* 2004), which uses a combination of a fully multivariate Taylor expansion and a Laplace approximation, and MLwiN (Rasbash *et al.* 2005), which implements quasi-likelihood algorithms (MQL and PQL). Finally, the econometric program LIMDEP (Greene 2007) uses simulated maximum likelihood.

Software for Bayesian estimation

Bayesian Markov Chain Monte Carlo (MCMC) algorithms are available in Mplus (Muthén and Muthén 2010) and MLwiN (Rasbash *et al.* 2005).

MCMC algorithms can be also implemented using the freeware BUGS (Spiegelhalter *et al.* 1997) and its Windows version WinBUGS (Lunn *et al.* 2000). Marshall and Spiegelhalter

(2001) provide an example of multilevel modelling using BUGS, including some syntax and discussion of the program.

19.4 Multilevel models for ordinal data in practice: an application to student ratings

In this section we present an application of multilevel models for ordinal responses to data on student satisfaction about university courses. We give guidelines on model specification, estimation and interpretation. The analysis is carried out with the `R` package `ordinal` which yields ML estimates using adaptive Gaussian quadrature (Christensen 2010). The dataset and the `R` script can be downloaded from the book web site.

Student ratings are an old and widely recognized instrument to evaluate university courses. The statistical analysis of student ratings calls for special techniques which take into account the ordinal nature of the response and the hierarchical structure of the phenomenon (ratings are nested in courses which are nested in schools). Moreover, in order to use the student ratings to measure the course quality, it should be recognized that the student satisfaction depends not only on the characteristics of the course (lecture hall, clarity of the teacher, textbook, and so on), but also on the traits and expectations of the student. Therefore, a fair comparison among courses requires the calculation of net measures adjusting for individual characteristics. To this end, multilevel modelling is a well suited technique (Grilli and Rampichini 2009).

For this application we use the data of Rampichini *et al.* (2004), which are gathered from a survey carried out by the University of Florence in the second semester of the academic year 1999/2000. The dataset is restricted to the courses with at least five respondents taken during the first year in the School of Engineering. The number of courses evaluated is 30 and the number of questionnaires is 767, while the number of questionnaires per course varies from 5 to 60. The items of the questionnaire require a response on the following 4-point ordinal scale: 1. *decidedly no*; 2. *more no than yes*; 3. *more yes than no*; 4. *decidedly yes*.

The main goal of the analysis is to identify 'good' and 'bad' courses on the basis of the student overall satisfaction about the course (`satisfaction`) while adjusting for student characteristics. In particular, we consider a binary variable for the full-time status (`fulltime`) and three self-assessed individual characteristics measured on the ordinal 4-point scale mentioned above: attendance with the intention of taking the exam in the first examination session (`exam`), previous knowledge of the subject (`knowledge`), and interest in the subject (`interest`).

The ordinal response `satisfaction` is studied via the random intercept cumulative model (19.4) using the logit link and C=4 categories:

$$logit(\gamma_{cij}) = \alpha_c - (\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j) \qquad c = 1, 2, 3,$$

where $j = 1, 2, \ldots, 30$ is the course index and i is the student index, while γ_{cij} is the cumulative probability up to the c-th category for student i in course j. The covariate vector $\mathbf{x}_{ij}$ includes the student characteristics, whereas the term u_j is a random effect representing unobserved factors at the course level interpretable as 'perceived quality'.

The analysis begins with the random intercept model without covariates (null model). This model is a benchmark for subsequent models and provides a cluster variance $\hat{\sigma}^2_u = 0.8800$ (the standard deviation is $\hat{\sigma}_u = 0.9381$). To test whether the cluster variance is statistically

significant, we compare the models with and without random effects. The LRT statistic is 98.12 with 1 *df* and a tiny p-value so that the null hypothesis is rejected (as noted in section 19.3.8, the p-value must be halved, even if in this case the result of the test is unchanged). Therefore, there is evidence of unobserved heterogeneity at course level: as expected, the courses have different levels of satisfaction.

The sample frequencies of the response (0.12, 0.27, 0.41, 0.20) are equal to the estimated probabilities from the single-level model (19.1) without covariates, which can be computed using the estimated thresholds ($\hat{\alpha}_1^\circ = -1.9685, \hat{\alpha}_2^\circ = -0.4480, \hat{\alpha}_3^\circ = 1.3572$). For example, the marginal probability that a student responds *more yes than no* ($Y_{ij} = 3$) is

$$\hat{\pi}_3^\circ = \hat{\gamma}_3^\circ - \hat{\gamma}_2^\circ = \frac{1}{1 + e^{-\hat{\alpha}_3^\circ}} - \frac{1}{1 + e^{-\hat{\alpha}_2^\circ}} = \frac{1}{1 + e^{-1.3572}} - \frac{1}{1 + e^{0.4480}} = 0.41.$$

A similar computation with the random intercept null model gives the conditional probabilities, i.e. the probabilities for a course with a hypothetical value of the random effect (see section 19.3.5). For example, given the estimated thresholds ($\hat{\alpha}_1 = -2.2397, \hat{\alpha}_2 = -0.5379, \hat{\alpha}_3 = 1.5624$), the probability that a student responds *more yes than no* ($Y_{ij} = 3$) for a course with a mean level of satisfaction ($u_j = 0$) is

$$\hat{\pi}_3 = \hat{\gamma}_3 - \hat{\gamma}_2 = \frac{1}{1 + e^{-\hat{\alpha}_3 + u_j}} - \frac{1}{1 + e^{-\hat{\alpha}_2 + u_j}} = \frac{1}{1 + e^{-1.5624}} - \frac{1}{1 + e^{0.5379}} = 0.46.$$

The amount of course-level unobserved heterogeneity is summarized by the ICC $\widehat{\rho} = 0.8800/(0.8800 + \pi^2/3) = 0.21$ (see section 19.3.3): this means that about one fifth of the total variability in the underlying satisfaction is at the course level. This is best appreciated by comparing some conditional probabilities as explained in section 19.3.3, for example $Pr(Y_{ij} \geq 3 \mid u_j = -1.96 \times 0.9381) = 0.21$ and $Pr(Y_{ij} \geq 3 \mid u_j = +1.96 \times 0.9381) = 0.92$: thus, the probability that a student rates a course positively (*more yes than no* or *decidedly yes*) ranges from 0.21 for a 'bad' course (at the 2.5th percentile) to 0.92 for a 'good' course (at the 97.5th percentile).

The analysis goes on by adding the covariates representing the characteristics of the students. Each of the variables measured on a 4-point ordinal scale (`exam`, `knowledge`, `interest`) is tried in two alternative codings: a set of 3 binary indicators, and a single numerical covariate with values -2, -1, 0 and 1 (the third category is thus taken as the baseline). The second coding, which is more parsimonious and easier to interpret, is chosen on the basis of the LRT.

The estimates for the random intercept cumulative model are reported in Table 19.1, along with the predicted conditional probabilities discussed in sections 19.3.3 and 19.3.5. In particular, the baseline student is a student who is not fulltime and who responds *more yes than no* to `exam`, `knowledge` and `interest`, while the baseline course has an average level of satisfaction, i.e. $u_j = 0$. The table shows how the baseline predicted probabilities change for a unit increase of each covariate and for a 'bad' course ($u_j = -1.96\sigma_u$) and a 'good' course ($u_j = 1.96\sigma_u$).

The effects of the student characteristics on the level of satisfaction are in the expected direction, i.e. the probability of being satisfied is higher for full-time students, for students intending to take the exam in the first examination session, for students with good background knowledge, and for students interested in the subject. The last feature has the largest effect, even if its estimate may be biased by endogeneity due to reverse causality.

Table 19.1 Estimates, standard errors and predicted probabilities $\hat{\pi}_c$ for the random intercept proportional odds model on the overall satisfaction for the course (University of Florence, School of Engineering, academic year 1999/2000)

	Estimate	*Std. Error*	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$
Thresholds						
First	-3.2567	0.2851				
Second	-0.9063	0.2556				
Third	1.9603	0.2664				
Baseline			0.04	0.25	0.59	0.12
Slopes						
Fulltime	0.3740	0.1808	0.03	0.19	0.61	0.17
Exam	0.4530	0.0901	0.02	0.18	0.61	0.18
Knowledge	0.5344	0.0882	0.02	0.17	0.61	0.19
Interest	1.2309	0.0966	0.01	0.09	0.57	0.33
Random effects						
Course-level σ_u	0.9477					
'Bad' course $(-1.96\sigma_u)$			0.20	0.53	0.26	0.02
'Good' course $(+1.96\sigma_u)$			0.01	0.05	0.47	0.47

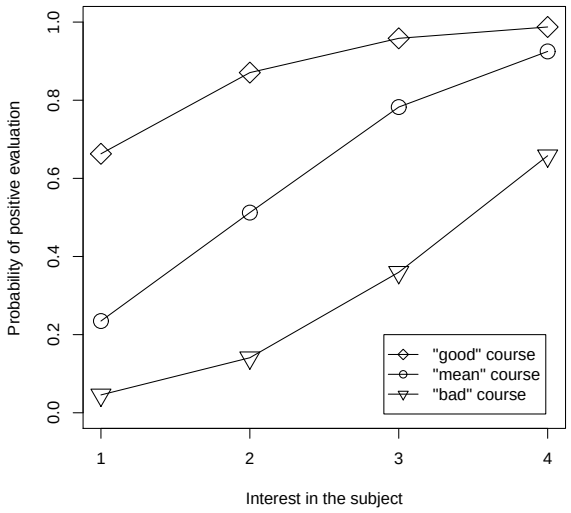

Figure 19.2 Probability of positive evaluation (*more yes than no* or *decidedly yes*) versus interest in the subject, for a full-time student responding *more yes than no* to the questions on exam and previous knowledge; the probabilities are conditional on three hypothetical values of the random effect ('good' course: $u_j = +1.96\sigma_u$; 'mean' course: $u_j = 0$; 'bad' course: $u_j = -1.96\sigma_u$).

The random effects represent the course-level unobserved heterogeneity in the ratings after controlling for the student characteristics: therefore, they may be interpreted as net measures of satisfaction for the course. The last two lines of Table 19.1 make clear that the courses are

quite different in terms of satisfaction and that the features of the course have an overall effect on the ratings higher than any of the features of the students (e.g. the baseline probability 0.12 of being very satisfied becomes 0.19 for a student with fully adequate previous knowledge and 0.47 for a student attending a 'good' course).

For an effective communication of the results, it is helpful to draw graphs of the predicted probabilities such as the one in Figure 19.2, where the probability of positive evaluation (*more yes than no* or *decidedly yes*) is plotted against the interest in the subject, for a full-time student responding *more yes than no* to the questions on exam and previous knowledge. The probabilities are computed for three hypothetical courses defined by fixing the random effect to 0 and $\pm 1.96\sigma_u$. The graph highlights that the effect of the interest in the subject is weak for good courses, which receive favorable evaluations anyway.

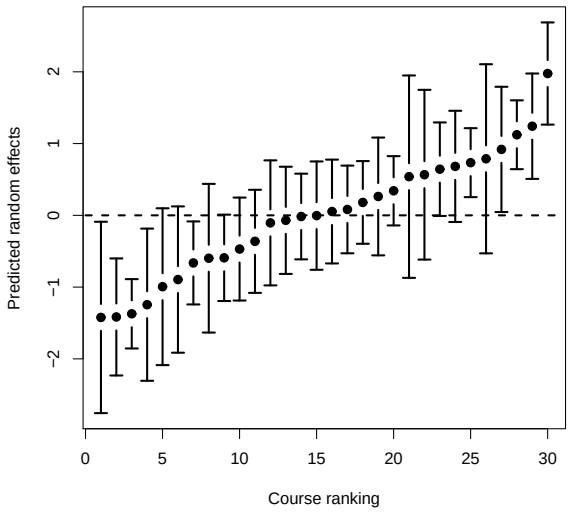

Figure 19.3 EB predictions of random effects with 95% confidence intervals.

Compared with the null model, the course-level standard deviation is nearly unchanged: this means that in the linear model for the underlying satisfaction (19.5), the reduction of the level 2 variance due to the covariates is similar to the reduction of the level 1 variance (see section 19.3.4). The course-level variance could be reduced by course-level covariates, such as the subject of the course or some features of the teacher. However, the dataset does not include course-level covariates, so the regression model can adjust the evaluations for the student characteristics, but it cannot explain why the adjusted evaluations are different among courses.

An effective way to report the course evaluations adjusted for the student characteristics is the 'caterpillar' graph in Figure 19.3, where the EB predicted random effects are plotted in ascending order along with 95% confidence intervals based on comparative standard errors (see section 19.3.9). Each confidence interval has a length inversely related to the number

of collected ratings for the course and it can be used to test whether the random effect of the corresponding course is significantly different from zero, which is the population mean: therefore, a course whose interval is entirely above (below) zero has an adjusted satisfaction significantly higher (lower) than the mean. In this application, it turns out that 10 courses have an adjusted satisfaction significantly different from the population mean (5 higher and 5 lower): such courses should be inspected to establish good and bad practices and to plan interventions for increasing the overall quality.

References

Agresti A 2010 *Analysis of Ordinal Categorical Data, 2nd edn*. Wiley, New York.

Agresti A and Natarajan R 2001 Modeling clustered ordered categorical data: A survey. *International Statistical Review* **69**, 345-371.

Austin PC 2010 Estimating Multilevel Logistic Regression Models When the Number of Clusters is Low: A Comparison of Different Statistical Software Procedures. *The International Journal of Biostatistics* **6**, Iss. 1, Art. 16.

Aitkin M 1999 A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–128.

Bauer DJ 2009 A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. *Psychometrika* **74**, 97–105.

Berkhof J and Snijders TAB 2001 Variance Component Testing in Multilevel Models. *Journal of Educational and Behavioral Statistics* **26**, 133–152.

Brant R 1990 Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* **46**, 1171–1178.

Breen R and Luijkx R 2010 Mixture Models for Ordinal Data. *Sociological Methods & Research* **39**, 3-24.

Browne WJ and Draper D 2006 A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **1**, 473–550.

Christensen RHB 2010 *ordinal: Regression Models for Ordinal Data*, sofware and manual downloadable from: http://cran.r-project.org/web/packages/ordinal/index.html

Cox C 1995 Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in medicine* **14**, 1191–1203.

Fielding A 1997 On scoring ordered classifications. *British Journal of Mathematical and Statistical Psychology* **50**, 285-307.

Fielding A 2004 Scaling for Residual Variance Components of Ordered Category Responses in Generalised Linear Mixed Multilevel Models. *Quality and Quantity* **38**, 425–433.

Greene WH 2007 *LIMDEP 9.0, Econometric Modeling Guide*. Econometric Software Inc., Plainview, NY.

Greene WH and Hensher DA (2010). *Modeling Ordered Choices: A Primer*. Cambridge UK: Cambridge University Press.

Grilli L and Rampichini C 2002 Specification issues in stratified variance component ordinal response models. *Statistical Modelling* **2**, 251–264.

Grilli L and Rampichini C 2009 Multilevel models for the evaluation of educational institutions: a review. In *Statistical Methods for the Evaluation of Educational Services and Quality of Products* (ed. Monari P, Bini M, Piccolo D and Salmaso L), pp. 61-80. Physica-Verlag, Heidelberg.

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. Springer, New York.

Hedeker D and Gibbons RD 1996 MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* **49**, 157–176.

Hedeker D and Gibbons RD 2008 *Supermix Mixed Effects Models*. Scientific Software International, Chicago.

Hilbe MH 2009. *Logistic Regression Models*. Chapman & Hall/CRC.

Johnson TR 2003 On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika* **68**, 563–583.

Johnson VE and Albert JH 1999. *Ordinal Data Modeling*. New York: Springer

Joe H 2008 Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis* **52**, 5066–5074.

Lillard LA and Panis CWA 2003 *aML Multilevel Multiprocess Statistical Software, Version 2.0.* EconWare, Los Angeles, CA.

Lindsay BG, Clogg CC and Grego J 1991 Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86**, 96–107.

Lunn DJ, Thomas A, Best N and Spiegelhalter D 2000 WinBUGS, a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.
Marshall EC and Spiegelhalter D 2001 Institutional performance. In: Leyland AH and Goldstein H, editors. *Multilevel Modelling of Health Statistics* 127-142. Wiley, New York.
Masters GN 1982 A Rasch model for partial credit scoring. *Psychometrika* **47**, 149-174.
McCullagh P 1980 Regression models for ordinal data. *Journal of the Royal Statistical Society Series B* **42**, 109-142.
McKelvey RD and Zavoina W 1975 A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* **4**, 103–120.
Mealli F and Rampichini C 1999 Estimating binary multilevel models through indirect inference, *Computational Statistics & Data Analysis* **29**, 313–324.
Muthén BO and Kaplan D 1985 A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology* **38**, 171–189.
Muthén BO and Satorra A 1995 Technical aspects of Muthns LISCOMP approach to estimation of latent variables relations with a comprehensive measurement model. *Psychometrika* **60**, 489–503.
Muthén LK and Muthén BO 2010. *Mplus Users Guide. Sixth Edition*. Muthén & Muthén, Los Angeles, CA.
O'Connell AA 2006. *Logistic Regression Models for Ordinal Response Variables*. Sage.
Olsson U 1979 On The Robustness Of Factor Analysis Against Crude Classification Of The Observations. *Multivariate Behavioral Research* **14**, 485–500.
Peterson B and Harrell FE 1990 Partial proportional odds models for ordinal response variables. *Applied Statistics* **39**, 205-217.
Pinheiro JC and Bates DM 1995 Approximations to the log-likelihood function in the nonlinear mixed effects model. *J. Computnl Graph. Statist.* **4**, 12-35.
Rabe-Hesketh S, Pickles A and Skrondal A 2003 Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling* **3**, 215–232.
Rabe-Hesketh S and Skrondal A 2008. *Multilevel and Longitudinal Modeling using Stata, 2nd edition*. Stata Press, College Station, TX.
Rabe-Hesketh S, Skrondal A and Pickles A 2005 Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* **128**, 301-323.
Moineddin R, Matheson FI and Glazier RH 2007 A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology* **7**:34.
Rampichini C, Grilli L and Petrucci A 2004 Analysis of University Course Evaluations: from descriptive measures to multilevel models *Statistical Methods & Applications* **13**, 357–373.
Rasbash J, Steele F, Browne WJ and Prosser B. *A Users Guide to MLwiN. Version 2.0*. Centre for Multilevel Modelling, University of Bristol, Bristol.
Raudenbush SW and Bryk AS 2002 *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd edn*. Sage.
Raudenbush SW, Bryk AS, Cheong YF and Congdon R. 2004 *HLM 6: Hierarchical Linear and Nonlinear Modeling. Scientific* Software International, Chicago.
Raudenbush SW , Yang ML and Yosef M. 2000 Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* **9**, 141-157.
Rodrguez G and Goldman N 1995 An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A* **158**, 73–89.
Rubin DB 1976 Inference and missing data. *Biometrika* **63**, 581–592.
Samejima F 1969 Estimation of Latent Trait Ability Using A Response Pattern of Graded Scores. *Psychometric Monograph* **17**, Psychometric Society, Bowling Green, OH.
SAS Institute 2009 *SAS/STAT(R) 9.2 User's Guide, Second Edition*. SAS Institute, Cary, NC.
Skrondal A and Rabe-Hesketh S 2004 *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC Press, Boca Raton, FL.
Skrondal A and Rabe-Hesketh S 2009 Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society Series A* **172**, 659-687.
Snijders TAB and Berkhof J 2008 Diagnostic Checks for Multilevel Models. In *Handbook of Multilevel Analysis* (ed. de Leeuw J and Meijer E) pp. 141–175. Springer, New York.
Snijders TAB. and Bosker RJ 1999. *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*. Sage, London.
Spiegelhalter DJ, Thomas A, Best NG and Gilks W 1997 *BUGS: Bayesian inference using Gibbs sampling (Version 0.60)*. Medical Research Council, Biostatistic Unit, Cambridge.
Stram DO and Lee JW 1994 Variance Component Testing in the Longitudinal Mixed Effects Model *Biometrics* **50**, 1171–1177.
Thissen D and Steinberg L 1986 A taxonomy of item response models. *Psychometrika* **51**, 567-577.
Verbeke G and Molenberghs G 2003 The Use of Score Tests for Inference on Variance Components. *Biometrics* **59**, 254-262.

Vermunt JK and Magidson J 2005 *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Statistical Innovations Inc., Belmont, MA.

Winship C and Mare RD 1984 Regression models with ordinal variables. *American Sociological Review* **49**, 512–525.