Revista Científica da Universidade José Eduardo dos Santos
e-ISSN: 3006-9688 | Vol. 05 | n.º 01 | 2025

EXTENSION OF PANEL DATA MODELS: RANDOM UTILITY MODELS FOR ORDERED CHOICES

EXTENSÃO DOS MODELOS COM DADOS EM PAINEL: MODELOS DE UTILIDADE ALEATÓRIA PARA ESCOLHAS ORDENADAS

Tadeu Fecayamale Leonardo [1]*; Gilmar da Conceição [2]*

[1] Faculdade de Economia da UJES. Huambo-Angola. [2] Faculdade de Economia e Gestão da Universidade Lincungo. Zambezia – Moçambique. * Email para correspondência: tad.eufeca@hotmail.com

ABSTRACT

Health satisfaction serves as a key indicator of perceived quality in health care services and is often used as a proxy for well-being in empirical research. This study explores the application of ordered probit models within the framework of random utility theory, particularly in scenarios where respondents' preferences regarding health satisfaction are ordinally ranked. Using Stata, we replicate and analyze the results presented in Tables 18.21 and 18.22 of Greene (2018), focusing on the signs and marginal effects of explanatory variables on reported health satisfaction. Our replication confirms the original findings for pooled models, traditional random effects models, and the Mundlak-adjusted random effects model. However, limitations emerged when attempting to replicate the conditional and unconditional fixed effects estimators, primarily due to data constraints and computational

RESUMO

A satisfação com a saúde constitui um indicador fundamental da qualidade percebida dos serviços de saúde e é frequentemente utilizada como proxy do bem-estar em pesquisas empíricas. Este estudo investiga a aplicação de modelos probit ordenados no contexto da teoria da utilidade aleatória, especialmente em situações nas quais as preferências dos respondentes em relação à satisfação com a saúde são classificadas ordinalmente. Utilizando o software Stata, replicamos e analisamos os resultados apresentados nas Tabelas 18.21 e 18.22 de Greene (2018), com ênfase nos sinais e efeitos marginais das variáveis explicativas sobre a satisfação autorreferida com a saúde. Nossa replicação confirma os resultados originais para os modelos agrupados (pooled), de efeitos aleatórios tradicionais e de efeitos aleatórios com o ajuste de Mundlak. Contudo, surgiram limitações na tentativa de replicar os estimadores de efeitos fixos condicionais e incondicionais, sobretudo devido a restrições dos dados e desafios

challenges. Nevertheless, we observe that the core differences across models lie mainly in coefficient magnitudes, while the direction (sign) of the estimated effects remains consistent. This suggests that individual heterogeneity influences the intensity—but not the direction—of health satisfaction responses.

Keywords: Ordered probit model, Health satisfaction, Panel data.

computacionais. Ainda assim, observamos que as principais diferenças entre os modelos residem na magnitude dos coeficientes, sendo que a direção (sinal) dos efeitos estimados permanece consistente. Isso sugere que a heterogeneidade individual influência a intensidade — mas não a direção — das respostas de satisfação com a saúde.

Palavras-chave: Modelo probit ordenado, Satisfação com a saúde, Dados em painel

1. INTRODUCTION

The identification and modeling of individual-level determinants of subjective well-being, particularly health satisfaction, have gained increasing attention in the field of applied microeconometrics. Health satisfaction is not only a key outcome of interest in public health and welfare analysis, but it also serves as a proxy for broader measures of individual utility, providing valuable insight for the design and evaluation of public policies. In this context, econometric techniques that appropriately handle ordinal dependent variables are essential for robust inference.

This paper focuses on the application and extension of random utility models for ordered choices, a class of models particularly well-suited for empirical settings where the outcome variable exhibits a natural ranking, but lacks cardinal interpretation. Examples include satisfaction ratings, self-reported health levels, and other survey-based opinion scales. Specifically, we employ the ordered probit model and its extensions to analyze the determinants of health satisfaction using panel data.

The present work was developed as part of the advanced econometrics module on extensions of panel data models, coordinated by Professor Miguel Portela in the second year of the PhD in Economics at the University of Minho. It builds upon theoretical and empirical foundations discussed in Greene (2018), with a particular focus on the estimation exercises presented in Tables 18.21 and 18.22. The main objective is to replicate and critically interpret these results while applying alternative panel specifications to account for unobserved heterogeneity.

In modeling latent preferences associated with health satisfaction, the ordered probit model treats the dependent variable as an ordinal indicator of an underlying, unobserved continuous utility. The probability of observing a given response category is expressed as a function of explanatory variables capturing both observable characteristics (such as income, education, and age) and unobserved individual-specific effects, which may persist over time (Wooldridge, 2009).

To this end, we estimate five variations of the ordered probit model using Stata:

1. Pooled Ordered Probit,
2. Traditional Random Effects Probit,
3. Mundlak-Adjusted Random Effects Probit,
4. Unconditional Fixed Effects Probit, and
5. Conditional Fixed Effects Probit (where estimation is computationally feasible).

These models differ in their treatment of unobserved heterogeneity. The Mundlak approach, in particular, augments the random effects specification by including the individual-specific means of time-varying regressors. This enables a partial correction for the endogeneity associated with omitted time-invariant variables, effectively creating a bridge between the fixed and random effects frameworks (Portela, 2023). Such an approach is especially useful when the assumption of strict exogeneity is tenuous.

Ordered response models, as noted by Greene and Hensher (2009), represent natural extensions of binary discrete choice models. They are widely used across applied economic fields—such as health economics, education, and finance—whenever the dependent variable reflects ranked outcomes (e.g., "poor", "fair", "good", "excellent"). Within the random utility framework, individuals are assumed to derive utility from latent continuous preferences, with observed ordinal responses reflecting the interval into which this latent utility falls.

Panel data introduces additional challenges and opportunities. In particular, correlation across time for the same individual necessitates model extensions that can control for unobserved heterogeneity, such as fixed and random effects structures (see Portela, 2023). These extensions

are crucial for obtaining unbiased and consistent estimates in the presence of intra-individual correlation and potential endogeneity.

In our empirical application, the main analytical focus lies on interpreting two sets of results:

- Table 18.21: Estimated coefficients (with emphasis on their sign and magnitude), and
- Table 18.22: Marginal effects of key explanatory variables on the probability of an individual reporting higher levels of health satisfaction.

By comparing the outcomes across different model specifications, we aim to assess the robustness of the estimated relationships and to highlight the econometric trade-offs involved in choosing among pooled, random effects, and fixed effects models for ordinal data.

This paper contributes to the literature by emphasizing the importance of model specification in the analysis of ordered outcomes in panel settings, particularly regarding subjective well-being measures. Moreover, it reinforces the practical utility of econometric replication exercises in enhancing methodological understanding and empirical rigor.

2- RANDOM UTILITY MODELS FOR ORDERED CHOICES

Random Utility Models (RUMs) for ordered choices constitute a foundational framework in the econometric analysis of discrete, ordinal outcomes. These models are widely employed to describe the data-generating process for variables that reflect individual preferences or evaluations on a naturally ordered categorical scale, without assuming cardinality. As noted by Greene and Hensher (2009), such models are particularly well-suited for analysing survey data where responses are captured in terms such as “strongly disagree” to “strongly agree”, or “very dissatisfied” to “very satisfied”.

Applications of ordered choice models span a broad array of disciplines, including economics, sociology, education, and public health. Notable empirical contributions include: bond ratings (Terza, 1985), credit ratings (Cheung, 1996), educational attainment (Machin and Vignoles, 2005), self-reported health status (Jones, Koolman, and Rice, 2003), job skill assessments (Marcus and Greene, 1985), and life satisfaction (Clark, Georgellis, and Sanfey, 2001). These examples highlight the versatility and relevance of ordered choice models in capturing subjective assessments in diverse socio-economic contexts.

According to Greene (2018), the core idea underpinning ordered response models is the existence of a latent continuous utility variable U^*_{ig}, which reflects an individual's unobserved strength of preference for a good or service g. The observed response R_{ig} is interpreted as a discretised representation of this latent utility, segmented by a set of threshold parameters μ_j, where $j = 1, 2, \ldots, j-1$. Formally, the model assumes:

$$-\infty < U^*_{ig} < +\infty, \tag{1}$$

$$R_{ig} = \begin{cases} 1 \ if \ -\infty < U^*_{ig} \leq \mu_1 \\ 2 \ if \ \mu_1 < U^*_{ig} \leq \mu_2 \\ 3 \ if \ \mu_2 < U^*_{ig} \leq \mu_3 \\ \quad . \\ \quad . \\ \quad . \\ J \ if \ \mu_{J-1} < U^*_{ig} < +\infty \end{cases}$$

Here, i indexes individuals, and g refers to the good or service being evaluated. The thresholds $\mu_1, \mu_2, \mu_3, \ldots, \mu_{J-1}$ are estimated parameters that partition the real line into J ordered categories. This formulation allows the model to reflect varying intensities of preference or satisfaction, as expressed through ordinal survey responses.

The latent utility U^*_{ig} is typically modeled as a linear function of observed covariates and a random error term:

$$U^*_{ig} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK} + \varepsilon_{ig}, \tag{2}$$

where X_{ik} represents the $k-th$ explanatory variable for individual i, β_K are the associated coefficients, and ε_{ig} is a random disturbance term. Assuming $\varepsilon_{ig} \sim N(0,1)$, the model becomes an ordered probit model, allowing for estimation via maximum likelihood techniques (See Verbeek, 2017). Accordingly, the normality assumption facilitates the derivation of response probabilities and ensures tractability of the likelihood function.

One of this model's conceptual strengths lies in its nonlinear mapping from latent utility to observed outcomes via the threshold structure. However, this is also a source of interpretive complexity. The intervals between ordinal categories are not assumed to be equidistant in utility space (Green, 2017). That is, the difference in utility between categories 2 and 3 may differ from that between categories 4 and 5. This feature implies a strictly nonlinear transformation, which is fully captured by the estimated threshold parameters (Verbeek, 2017).

Nonetheless, one limitation of the standard ordered probit model is that it treats the thresholds as fixed across individuals. This may be problematic if there is unobserved heterogeneity in how individuals perceive or use the response scale (Green, 2017). Moreover, the assumption that the covariates enter the utility function linearly and with homogeneous effects across thresholds may not always hold in practice.

Extensions such as random effects ordered probit models or Mundlak-adjusted specifications are employed to address such limitations, especially in the context of panel data. These allow for individual-specific unobserved effects and help control for time-invariant omitted variables, improving the model's ability to capture persistent differences in preferences or reporting behavior (Verbeek, 2017).

2.1 Econometric Framework

We begin by modeling individual health satisfaction using an **ordered probit structure,** grounded in the random utility model. Let Y_{it}^{*} denote the unobserved latent variable representing individual i's utility or satisfaction level at time t, such that:

$$Y^{*}{}_{it} = X_{it}{}^{T}\beta + \alpha_i + \varepsilon_{it}, \tag{3}$$

where:

- $X_{it}{}^{T}$ is a vector of time-varying observed characteristics (e.g., income, age, education),
- β is a vector of parameters to be estimated,
- α_i captures unobserved, individual-specific heterogeneity (random or fixed effects),
- $\varepsilon_{ig} \sim N(0,1)$ is an idiosyncratic error term.

The observed variable Y_{it}, representing the ordinal health satisfaction score, is derived from the latent variable via threshold cut-points:

$$Y_{it} = J \text{ if } \mu_{j-1} < Y^*_{it} \leq \mu_j, \quad (4)$$

with $\mu_j - \infty$ and $\mu_j + \infty$, and the μ_j are thresholds to be estimated.

2.2 Estimation Strategy

To evaluate the sensitivity of results to assumptions about unobserved heterogeneity, we estimate and compare five ordered probit model specifications:

(i) Pooled Ordered Probit Model

This baseline model ignores the panel structure and treats all observations as independent across individuals and time. It does not control for unobserved individual effects, potentially leading to biased estimates if relevant omitted variables are correlated with the regressors (Greene, 2018).

(ii) Traditional Random Effects Ordered Probit

In this specification, $\alpha_i \sim N(0, \sigma_\alpha^2)$ is assumed to be uncorrelated with X_{it}. The likelihood function integrates over the distribution of the random effects using simulated maximum likelihood. This model accounts for intra-individual correlation over time but rests on the strict exogeneity assumption (Verbeek, 2017).

(iii) Mundlak-Adjusted Random Effects Model

To relax the orthogonality assumption of the random effects model, we estimate the Mundlak-adjusted model (Mundlak, 1978), which augments the regressors with the individual-specific means of time-varying covariates:

$$Y^*_{it} = X_{it}^T \beta + \bar{X}_i^T \delta + \alpha_i + \varepsilon_{it}, \quad (5)$$

Where $\bar{X}_i$ denotes the time-averaged covariates for individual i. This specification nests the fixed effects model and serves as a diagnostic tool to test for endogeneity between X_{it} and α_i.

(iv) Unconditional Fixed Effects Model

This approach attempts to estimate fixed effects directly without conditioning on sufficient statistics, which in nonlinear models is known to introduce incidental parameters bias (Neyman and Scott, 1948). In practice, this approach is often unstable, especially with short panels, and can lead to inconsistent estimates of the structural parameters.

(v) Conditional Fixed Effects Model

The conditional fixed effects ordered probit model seeks to eliminate individual effects by conditioning on a sufficient statistic. However, as noted by Greene (2018), this method is only feasible for certain nonlinear models, and its implementation in the context of ordered outcomes is often limited or inapplicable due to computational complexity and the lack of a sufficient statistic.

2.3 Marginal Effects

In addition to estimating the signs and magnitudes of the coefficients, we compute the marginal effects associated with the ordered probit models, specifically for the pooled and random effects specifications, following the procedure outlined in Table 18.22 of Greene (2018). Marginal effects provide an intuitive interpretation by quantifying the change in the probability of selecting a particular category of the dependent variable in response to a small change in a continuous regressor, or a discrete shift in a binary regressor.

Formally, the marginal effect of regressor x_k on the probability of observing outcome j is given by:

$$\frac{\partial \Pr(y_{it}=j|(X_{it})}{\partial x_{it,k}} = [\emptyset(\mu_{j-1} - X_{it}{}^{T}\beta) - \emptyset(\mu_j - X_{it}{}^{T}\beta)].\beta_K, \qquad (6)$$

where ∅ denotes the standard normal probability density function.

Moreover, these effects are typically evaluated at the sample means or for representative individuals. This step is crucial for policy analysis, as it facilitates a clearer understanding of how variations in individual characteristics influence the likelihood of different satisfaction levels. Consequently, marginal effects enhance the interpretability of the estimated model, bridging the gap between statistical output and economic insight.

2.4 Software Implementation

All estimations were performed using Stata, leveraging the built-in procedures for ordered probit models and user-written commands for panel data extensions. Special care was taken to match model specifications with those in Greene (2018), ensuring fidelity in the replication and comparability of results.

3- REPLICATION OF TABLE 18.21 FROM GREEN (2018)

Table 1: Estimated Ordered Probit Model for Health Satisfaction

HSAT	Ordered Probit Pooled	Ordered Probit RE	Ordered Probit RE Mundlak
Age	-0.0191***	-0.0340***	-0.0618***
	(0.0010)	(0.0013)	(0.0027)
Income	0.1812***	0.1060*	0.2777***
	(0.0512)	(0.0593)	(0.0696)
Kids	0.0608***	0.0134	0.0124
	(0.0211)	(0.0241)	(0.0223)
Education	0.0342***	0.0453***	0.0187
	(0.0044)	(0.0059)	(0.0262)
Married	0.0258	0.0784***	0.0255
	(0.0256)	(0.0298)	(0.0393)
Working	0.1293***	0.0643***	-0.0214
	(0.0211)	(0.0244)	(0.0272)
Averaged Age			0.0359***
			(0.0029)
Averaged Income			0.1303
			(0.1198)
Averaged Education			0.0258
			(0.0269)

Averaged Married			0.0300
			(0.0520)
Averaged Working			0.2531***
			(0.0447)
Observations	27326	27326	27326

Notes: standard errors in parentheses. Significance levels: *, 10%; **, 5%; ***, 1%.

Source: Author's computations based on replication of Greene (2018), Table 18.21

4- INTERPRETATION OF RESULTS

The results obtained from the pooled ordered probit model indicate that, on average, an increase in age is associated with a lower probability of reporting higher levels of health satisfaction. Conversely, higher income and more years of education are positively associated with the likelihood of higher satisfaction levels. Moreover, dummy variables for having children, being married, and being employed suggest that these characteristics are linked to higher reported health satisfaction, although not all coefficients are statistically significant across models.

However, the pooled model relies on the assumption that unobserved individual heterogeneity is uncorrelated with the explanatory variables. This limitation motivates the use of panel data techniques that explicitly account for such heterogeneity. While fixed effects ordered probit models (both conditional and unconditional) are better suited for this purpose, we faced software constraints that prevented their implementation in Stata. Nevertheless, Greene (2018) provides estimates from these models showing that, once unobserved heterogeneity is considered, the signs and significance of most covariates remain consistent with the pooled model, with the notable exception of the kids and working variables. In these cases, the sign of the marginal effect reverses, implying that individuals with children or those who are employed are less likely to report high health satisfaction when individual-specific effects are considered.

The random effects ordered probit model addresses heterogeneity differently. Unlike the pooled model, it assumes that unobserved heterogeneity is randomly distributed and uncorrelated with regressors. In our replication, the coefficient patterns are broadly like the pooled model, though some magnitudes differ. Importantly, the Mundlak-adjusted random effects specification introduces the time averages of time-varying covariates as additional regressors. This

modification relaxes the strict exogeneity assumption of the traditional random effects model and allows for correlation between individual effects and explanatory variables.

Our results from the Mundlak model are consistent with those reported by Greene (2018) and match closely with the fixed effects estimators in terms of signs and statistical significance. In particular, the working variable shows a negative association with health satisfaction when controlling for unobserved heterogeneity, in line with fixed effects estimates.

4.1- Partial Effects

To better understand the substantive implications of these models, we compute partial (marginal) effects for selected covariates. For example, the marginal effect of age indicates that a one-year increase raises the probability of reporting lower levels of health satisfaction across most categories. Specifically, for satisfaction levels 0 through 7, the probability increases by 0.0061, 0.0003, 0.0008, 0.0012, 0.0012, 0.0024, 0.0008, and 0.0008, respectively. Meanwhile, the probability of reporting the highest satisfaction levels (8, 9, and 10) decreases by 0.0019, 0.0021, and 0.0035, respectively, holding all other variables constant.

Similarly, an increase in income reduces the probability of reporting lower satisfaction categories—for instance, by 0.0061 (level 0), 0.003 (1), 0.0072 (2), and 0.0113 (3)—and increases the probability of reporting higher satisfaction levels (e.g., 0.0184 at level 8, 0.0198 at level 9, and 0.0336 at level 10). Education has comparable effects: one additional year of schooling decreases the likelihood of selecting lower satisfaction categories and increases the probability of reporting high satisfaction from category 8 onward.

Dummy variables such as married and working follow a similar pattern, although their interpretation is less straightforward due to their binary nature.

5- CONCLUSIONS

The conclusions derived from this analysis fall into two main categories: substantive implications for the determinants of health satisfaction, and methodological insights regarding the choice of econometric model.

From a substantive perspective, health satisfaction is significantly influenced by individual characteristics, particularly age, income, education, and family and employment status. Failing to account for these factors can lead to biased or misleading conclusions. Moreover, the scale used in ordered response models implies that the strength of preference between adjacent categories is not constant — i.e., the difference in utility between categories 1 and 2 may not equal that between categories 3 and 4.

From a methodological perspective, our replication confirms that pooled and random effects ordered probit models produce qualitatively similar results to those reported in Greene (2018). However, only the Mundlak correction adequately controls for individual heterogeneity in the presence of potentially endogenous regressors. While we were unable to estimate fixed effects ordered probit models in Stata—possibly due to the large number of dummy variables in the dataset—the consistency of the Mundlak estimates with Greene's fixed effects results provides reassurance about the robustness of our findings.

In sum, this replication highlights the importance of accounting for unobserved heterogeneity in ordered response models and demonstrates the practical usefulness of the Mundlak correction as a middle ground between pooled and fixed effects specifications.

REFERENCES

Greene, W. H. (2017). *Econometric analysis* (8th ed.). Pearson.

Greene, W. H. (2018). *Econometric analysis* (8th ed.). Pearson.

Greene, W. H., & Hensher, D. A. (2009). *Modeling ordered choices* (Unpublished manuscript), 1–181.

Greene, W. H., & Hensher, D. A. (2009). *Ordered choices and heterogeneity in attribute processing. Journal of Transport Economics and Policy*, forthcoming. https://doi.org/10.2307/25801404

Portela, M. (2023). *Panel data: A helicopter tour* [Lecture notes].

Verbeek, M. (2017). *A guide to modern econometrics* (5th ed.). John Wiley & Sons.

Winkelmann, R. (2005). Subjective well-being and the family: Results from an ordered probit model with multiple random effects. *Empirical Economics, 30*(3), 749–761. https://doi.org/10.1007/s00181-005-0255-7

Wooldridge, J. M. (2009). On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, 104(3), 112–114. https://doi.org/10.1016/j.econlet.2009.03.010