Journal of Economics and Econometrics Vol. 54, No.1, 2011 pp. 7-23

ISSN 2032-9652 E-ISSN 2032-9660

Estimating Ordered Categorical Variables Using Panel Data: A Generalised Ordered Probit Model with an Autofit Procedure

CHRISTIAN PFARR[‡], ANDREAS SCHMID[†] AND UDO SCHNEIDER[*]

ABSTRACT

Estimation procedures for ordered categories usually assume that the estimated coefficients of independent variables do not vary between the categories (parallel-lines assumption). This view neglects possible heterogeneous effects of some explaining factors. This paper describes the use of an autofit option for identifying variables that meet the parallel-lines assumption when estimating random-effects generalised ordered probit models. We combine the test procedure developed by Richard Williams (gologit2) with the random-effects estimation command regoprob by Stefan Boes.

JEL Classification: C23, C25, C87, I10.
Keywords: Generalised ordered probit, panel data, autofit, self-assessed health.

‡Christian Pfarr, University of Bayreuth, Department of Law and Economics, Institute of Public Finance, D-95447 Bayreuth, Germany. E-mail: christian.pfarr@uni-bayreuth.de, corresponding author.

†Andreas Schmid, University of Bayreuth, Department of Law and Economics, Institute of Public Finance, D-95447 Bayreuth, Germany. E-mail: andreas.schmid@uni-bayreuth.de.

*Udo Schneider, University of Bayreuth, Department of Law and Economics, Institute of Public Finance, D-95447 Bayreuth. E-mail: udo.schneider@uni-bayreuth.de.

1 INTRODUCTION

When estimating a model for ordered categorical variables, normally, one faces an all-or-nothing situation. On the one hand, estimation procedures for ordered categories usually assume that the estimated coefficients of independent variables do not vary between the categories (parallel-lines assumption, cf. Long 1997). This view neglects possible heterogeneous effects of some explaining factors. For example, the traditional ordered probit model implies that all variables are constrained and meet the parallel-lines assumption. On the other hand, a fully flexible approach (generalised ordered probit) allows all coefficients to vary across the categories, which again is a very strong assumption. Of course, manually setting only some variables as constrained would be an option. However, in most cases theory does not provide adequate guidance to determine those variables that do not vary. Thus, a pragmatic and empirically robust approach is wanted.

In contrast to cross-section data for which the procedure gologit2 (cf. Williams 2006) provides an automated selection mechanism, up to now, such an instrument was not available for panel data. Regoprob2, the STATA module proposed in this paper, presents a solution to this problem. It is a user-written program and an extension of regoprob that estimates random-effects generalised ordered probit models for ordinal dependent variables. It includes an optional automated fitting procedure for identifying the relevant variables that meet the parallel-lines assumption (cf. Pfarr, Schmid and Schneider 2010).

In the following we give a brief introduction to the theoretic background and illustrate the application and the benefits of regoprob2 using an estimation of self-assessed health.

2 CONCEPTUAL FRAMEWORK

When analysing ordered choice models, the presence or absence of individual heterogeneity is highly relevant. For instance, considering homogenous groups like “fruit flies” the assumptions of zero mean, homoscedasticity and homogenous thresholds are plausible without a doubt. However, the analysis of a population of individuals e.g. regarding their subjective well-being or self-assessed health status might be more complicated (cf. Greene and Hensher 2010, p. 208). The regression equation of an ordered categorical variable such as self-

assessed health (SAH) will include socio-economic variables like income, education, marital status or health related variables as well as a series of measurable and immeasurable factors affecting the decision to choose one of the health categories. This raises the question if a zero mean and homoscedastic errors can be presumed and if so, whether these assumptions can capture the existing heterogeneity adequately. Hence, the hypothesis of equal thresholds for all individuals is at least questionable (Greene and Hensher 2010).

More formally, consider the observed categorical variable self-assessed health with an underlying latent health status of the respondent y^*. In this case, ordered response models are the basic standard estimation procedure. Following the work of Boes and Winkelmann (2006) and focusing on the cross-section case first, let y be the ordered categorical outcome, $y \in \{1, 2,..., J\}$ where J denotes the number of distinct categories. The cumulative probabilities of the discrete outcome are then related to a set of explanatory variables x:

$$\Pr\left[y \leq j \mid x\right] = F\left(\kappa_j - x'\beta\right) \qquad j = 1,...,J \tag{1.1}$$

Here, $\boldsymbol{\kappa}_j$ are the unknown threshold parameters and $\boldsymbol{\beta}$ are the unknown coefficients.[1] The function F usually represents an accumulative standard normal or logistic distribution, resulting in an ordered probit model or an ordered logit model respectively. Including the underlying latent variable, this results in:

$$y = j \quad \text{if and only if} \quad \kappa_{j-1} \leq y^* = x'\beta + u < \kappa_j \; j = 1,...,J \tag{1.2}$$

This means that the thresholds divide the linear slope (y^*) into J categories. Moreover, observable and unobservable factors influence the latent variable health. For the latter factors, a zero mean and a constant variance is assumed, e.g. $\sigma^2 = 1$ for the ordered probit model.

The probability that a respondent reports his health status to be in category j can then be written as:

$$\Pr\left[y = j \mid x\right] = F\left(\kappa_j - x'\beta\right) - F\left(\kappa_{j-1} - x'\beta\right) \tag{1.3}$$

[1]One assumption on the threshold parameters is that $\kappa_j > \kappa_{j+1}, \forall j$ and that $\kappa_J = \infty$ and $\kappa_o - \infty$.

For identification purposes, it is necessary to set the constant of the regression to zero and to assume a constant variance.

However, one obstacle to the appropriate implementation of an ordered probit model is the single index or parallel-lines assumption (Long 1997). In traditional models for categorical dependent variables the coefficient vector β is assumed to be the same for all categories J. This means that with the increase of an independent variable, the cumulated distribution shifts to the right or left but there is no shift in the slope of the distribution. Boes and Winkelmann (2006), Greene et al. (2008) and Pudney and Shields (2000) suggest that in the set of thresholds, individual variation is an indicator for heterogeneity that appears in the data and that this case is not reflected in traditional ordered probit models. Relaxing the assumption of equal thresholds for all individuals and allowing the indices to differ across the outcomes leads to a generalised ordered probit model. Here, the threshold parameters are individual specific and depend on the covariates:[2]

$$\kappa_{ij} = \tilde{\kappa}_j + x_i'\gamma_j, \tag{1.4}$$

where γ_j are the influence parameters of the covariates on the thresholds. Entering the threshold equation (1.4) into the cumulative probability of the generalised ordered probit model leads to the following expression:

$$\Pr\left[y \leq j \mid x\right] = F\left(\tilde{\kappa}_j + x_i'\gamma_j - x_i'\beta\right) = F\left(\tilde{\kappa}_j - x_i'\beta_j\right) \quad j = 1,\ldots,J \tag{1.5}$$

As one can see from equation (1.5), the coefficients of the covariates and the threshold coefficients cannot be identified separately when the same set of variables x is used. It follows that $\beta_j = \beta - \gamma_j$ and that the generalised ordered probit model has one index $x_i'\beta_j$ for each category j of the outcome variable.[3] This approach leads to the estimation of J-1 binary probit models (Williams 2006). The first model estimates category 1 versus categories 2,..., J; the second model does the same regarding categories 1 and 2 versus 3,..., J. Equation J-1 then compares the choice between categories 1,..., J-1 versus category J. This

[2]The predicted probabilities have to be in the (0;1) interval to fulfill the order condition in the generalised ordered probit model.

[3]The generalised ordered probit model nests the standard ordered probit model with the restriction that $\beta_1 = \ldots = \beta_{J-1}$.

specification allows for individual heterogeneity in the β-parameters that leads to heterogeneity across the categories of the dependent variable.

For panel data, individual heterogeneity is accounted for using a random-effects generalised ordered probit approach (cf. Boes 2007, p. 133). More formally, let SAH be an ordinal variable which takes on the values $j = 1,\ldots,J$. In contrast to the cross-section representation, the outcome probabilities are conditional on the individual effect α_i:[4]

$$\begin{aligned} &\Pr\left(Y_{it} = 1 \mid x_{it}, \alpha_i\right) = F\left(-x_{it}'\beta_1 - \alpha_i\right) \\ &\Pr\left(Y_{it} = y \mid x_{it}, \alpha_i\right) = F\left(-x_{it}'\beta_y - \alpha_i\right) - F\left(-x_{it}'\beta_{y-1} - \alpha_i\right) \qquad (1.6) \\ &\Pr\left(Y_{it} = J \mid x_{it}, \alpha_i\right) = 1 - F\left(-x_{it}'\beta_{J-1} - \alpha_i\right) \qquad j = 2,\ldots,J-1 \end{aligned}$$

For the individual effects, a zero mean and a constant variance σ^2 is assumed so that $\rho = \sigma^2 / (1 + \sigma^2)$. As for the cross-section version of the generalised ordered probit model, the approach allows any number of the βy (from none to all) to vary across the categories. Hence, using panel data allows for the inclusion of two kinds of heterogeneity. First, unobserved individual heterogeneity is captured by a random-effects specification. Second, differences in the cut-points and therefore in the beta coefficients represent the observed heterogeneity in the reporting of the categorical variable.

However, the problem of identifying the constrained variables remains unsolved. As pointed out above, theory often does not provide good guidance. As both extremes – setting all or none variables constrained – are equally unlikely, a pragmatic and empirically robust approach is needed. Building on the automated fitting procedure that Williams (2006) developed for gologit2 we suggest an iterative fitting process that we have implemented in regoprob2. The autofit option of regoprob2 triggers an iterative process used to identify the random-effects generalised ordered probit model that best fits the data.

At the beginning, an unconstrained model (all coefficients could vary) is estimated. Then, in a first step, a Wald test is applied on each variable to prove whether the coefficients differ across equations. The least significant variable is then set as constrained, that means to have equal effects over all categories. With autofit2 (*alpha*) one can choose

[4]Note that in equation (1.6) the beta coefficients differ between the categories of the dependent variable.

another significance level than the standard one. The parameter alpha is the desired significance level for the tests; alpha must be greater than 0 and less than 1. If autofit is specified without parameters, as in this case, the default alpha-value is .05. Note that the higher alpha is, the easier it is to reject the parallel lines assumption, and the less parsimonious the model will tend to be.[5] Then the model is refitted with the constraints identified so far and the step is repeated until only significant variables remain. Finally, as specification test, a global Wald test on the full model with constraints is applied to confirm the null hypothesis that the parallel-lines assumption is not violated. The following example illustrates the process and describes the fitting procedure in more detail.

3 ESTIMATING A GENERALISED ORDERED PROBIT MODEL WITH THE AUTOFIT OPTION: AN EXAMPLE

To discuss the estimation of random-effects generalised ordered probit models for ordered categorical variables we use self-assessed health as dependent variable (for variable description see table A1 in the Appendix). It is a 5-point categorical variable with 1 indicating very bad and 5 very good self-reported health status. As explanatory variables, a set of ten dummy variables indicating various diseases is used.[6] For illustration purposes, we restrict the analysis to a 10 %-random sample of the original SAVE data[7] consisting of 1,186 individuals for the years 2006 to 2008.

First, we start with a fully constrained model (random-effects ordered probit) (cf. Frechette 2001). As it is clear from the results presented below (see table 1), with the exception of gastric_ulcer, all other disease variables show the expected significant sign. The magnitude of the partial effects varies between the variables.

[5]This option may be time consuming depending on the sample size and the number of explanatory variables.

[6]For more details regarding reporting heterogeneity in self-assessed health see Schneider et al. (2011).

[7]The SAVE study is conducted by the Mannheim Research Institute for the Economics of Aging (MEA) and was started in 2001. Originally, the longitudinal study on households' financial behaviour focused on savings and old-age provisions but also deals with aspects of health and health behaviour (cf. Börsch-Supan et al. 2008).

Table 1: Results of the fully constrained random-effects ordered probit model

Random-Effects Generalised Ordered Probit				Number of obs	= 1186
				Wald chi2 (19)	= 415.84
Log likelihood = -1176.8221				Prob>chi2	= 0.0000
sah	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
mleq1					
backache	-1.0990	0.1295	-8.49	0.000	-1.35287 -0.84515
blood	-0.4476	0.1046	-4.28	0.000	-0.65272 -0.24257
cancer	-0.6491	0.2575	-2.52	0.012	-1.15377 -0.14441
chol	-0.3641	0.1257	-2.90	0.004	-0.61047 -0.11773
gastric_ulcer	-0.4359	0.2758	-1.58	0.114	-0.97654 0.10477
heart	-0.8273	0.1608	-5.14	0.000	-1.14259 -0.51210
mental	-0.5862	0.1809	-3.24	0.001	-0.94072 -0.23164
other_disease	-1.2175	0.1248	-9.75	0.000	-1.46211 -0.97279
pul_asthma	-0.8595	0.1911	-4.50	0.000	-1.23413 -0.48489
stroke	-0.7893	0.2676	-2.95	0.003	-1.31382 -0.26487
cut1 _cons	-4.7037	0.3560	-13.21	0.000	-5.40154 -4.00590
cut2 _cons	-3.2809	0.2417	-13.58	0.000	-3.75454 -2.80722
cut3 _cons	-1.1596	0.1053	-11.02	0.000	-1.36593 -0.95331
cut4 _cons	1.3583	0.1317	10.32	0.000	1.10018 1.61633
rho _cons	0.4632	0.0776	5.97	0.000	0.31119 0.61519

In contrast to the results above, a generalised ordered probit model allows different parameter vectors for each outcome. This means that we aim at assessing the observable individual heterogeneity in the threshold parameters as well as in the mean of the regression (cf. Greene and Hensher 2010). From table 2, it is obvious that the magnitude of the coefficients as well as the level of significance vary between the four binary probit models. The coefficients of backache are

significant throughout the equations and range from -0.66 to -1.52. While the ordered probit estimation shows a highly significant impact, the generalised model also implies an increasing significant negative coefficient. This means that individuals suffering from chronic backache are less likely to report a better health status. The effect is lower when comparing SAH categories 1 vs. 2-5, and highest for categories 1-4 vs. 5. For the variable blood, only equations 3 and 4 show a significant impact. People with hypertension tend to report the extreme categories of SAH less often. In consequence, those individuals will choose the middle categories more often. For heart diseases, it is obvious that there exists a tendency to assign oneself into the lowest categories of SAH.

If one looks at the overall significance reported by a likelihood ratio test, the generalised ordered probit model fails to reject the hypothesis that all coefficients have no influence. Consequently, a model with full variation seems to be overspecified and therefore unsuitable for estimating ordered categorical models.

Table 2: Random-effects generalised ordered probit with all variables varying

Random-Effects Generalised Ordered Probit					Number of obs =	1186
					Wald chi2(19) =	22.08
Log likelihood = -1145.8067					Prob>chi2 =	0.9904
sah	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mleq1						
backache	-0.9737	0.3137	-3.10	0.002	-1.58859	-0.35882
blood	0.0816	0.3133	0.26	0.794	-0.53241	0.69567
cancer	-0.2652	0.6970	-0.38	0.704	-1.63120	1.10090
chol	-0.4152	0.3221	-1.29	0.197	-1.04650	0.21607
gastric_ulcer	-0.2362	0.8399	-0.28	0.779	-1.88242	1.40995
heart	-0.7720	0.3364	-2.30	0.022	-1.43130	-0.11274
mental	-0.8017	0.3649	-2.20	0.028	-1.51683	-0.08652
other_disease	-1.1540	0.3172	-3.64	0.000	-1.77565	-0.53238
pul_asthma	-0.9270	0.4122	-2.25	0.024	-1.73484	-0.11922
stroke	-0.2663	0.6011	-0.44	0.658	-1.44450	0.91195

_cons	4.4089	0.4941	8.92	0.000	3.44061	5.37727
mleq2						
backache	-0.6614	0.1735	-3.81	0.000	-1.00143	-0.32140
blood	-0.1546	0.1656	-0.93	0.351	-0.47918	0.16999
cancer	-0.9161	0.3509	-2.61	0.009	-1.60380	-0.22835
chol	-0.1535	0.1928	-0.80	0.426	-0.53139	0.22439
gastric_ulcer	-0.0508	0.4014	-0.13	0.899	-0.83754	0.73587
heart	-0.8607	0.2173	-3.96	0.000	-1.28648	-0.43484
mental	-0.6308	0.2438	-2.59	0.010	-1.10859	-0.15293
other_disease	-0.9808	0.1695	-5.79	0.000	-1.31308	-0.64861
pul_asthma	-1.0942	0.2605	-4.20	0.000	-1.60476	-0.58372
stroke	-1.0172	0.3393	-3.00	0.003	-1.68222	-0.35223
_cons	2.8767	0.2663	10.80	0.000	2.35469	3.39873
mleq3						
backache	-1.4291	0.1764	-8.10	0.000	-1.77489	-1.08332
blood	-0.6776	0.1364	-4.97	0.000	-0.94508	-0.41022
cancer	-0.4146	0.3314	-1.25	0.211	-1.06411	0.23494
chol	-0.4047	0.1642	-2.47	0.014	-0.72646	-0.08299
gastric_ulcer	-0.6336	0.4050	-1.56	0.118	-1.42742	0.16018
heart	-1.1488	0.2301	-4.99	0.000	-1.59976	-0.69787
mental	-0.5660	0.2467	-2.29	0.022	-1.04951	-0.08252
other_disease	-1.4553	0.1642	-8.86	0.000	-1.77708	-1.13344
pul_asthma	-0.7739	0.2395	-3.23	0.001	-1.24335	-0.30437
stroke	-0.7298	0.3659	-1.99	0.046	-1.44685	-0.01269
_cons	1.3808	0.1344	10.28	0.000	1.11746	1.64422
mleq4						
backache	-1.5165	0.4174	-3.63	0.000	-2.33458	-0.69839
blood	-0.4197	0.2088	-2.01	0.044	-0.82886	-0.01052
cancer	-6.0224	387.4540	-0.02	0.988	-765.41820	753.37340
chol	-0.7821	0.3335	-2.35	0.019	-1.43583	-0.12845

gastric_ulcer	-6.2791	430.5480	-0.01	0.988	-850.13760	837.57940
heart	-0.4606	0.4016	-1.15	0.251	-1.24767	0.32653
mental	-0.7315	0.6360	-1.15	0.250	-1.97810	0.51501
other_disease	-0.8873	0.2581	-3.44	0.001	-1.39320	-0.38135
pul_asthma	0.0785	0.4373	0.18	0.857	-0.77848	0.93556
stroke	-5.7461	546.0777	-0.01	0.992	-1076.0390	1064.54700
_cons	-1.3697	0.1539	-8.90	0.000	-1.67140	-1.06804
rho						
_cons	0.4824	0.0846	5.70	0.000	0.31665	0.64819

Thus, at this point, it has to be decided, which variables are most likely constrained and which should be allowed to vary. To the best knowledge of the authors, there is no good theory that would reliably predict if a certain illness presents a constrained or an unconstrained factor regarding SAH – a typical problem encountered in many similar cases. For this reason, we now apply the autofit procedure as suggested above.[8]

In our example, the first step in the estimation process is a model with full variation of all ten explanatory variables. After estimation of this model and Wald tests on each coefficient, the variable mental with a P-value of 0.9437 is identified as the least significant variable after the first step. Next, this procedure is repeated with the variable mental set as constrained. In step two, gastric_ulcer meets the parallel-lines assumption.

Table 3: An example of the autofit procedure

Testing the parallel lines assumption using the .05 level of significance...
Step 1: Constraints for parallel lines imposed for mental (P Value = 0.9437)
Step 2: Constraints for parallel lines imposed for gastric_ulcer (P Value = 0.7481)
Step 3: Constraints for parallel lines imposed for stroke (P Value = 0.6501)
Step 4: Constraints for parallel lines imposed for cancer (P Value = 0.5687)
Step 5: Constraints for parallel lines imposed for chol (P Value = 0.4278)

[8]For a more detailed discussion of the autofitting procedure see Williams, R. (2006) and for the theoretical background of estimating random-effects generalised ordered probit models see Boes (2007).

Step 6: Constraints for parallel lines imposed for heart (P Value = 0.2303)
Step 7: Constraints for parallel lines imposed for pul_asthma(P Value = 0.1287)
Step 8: Constraints for parallel lines are not imposed for
backache (P Value = 0.00156)
blood (P Value = 0.00332)
other_disease (P Value = 0.01315)

As can be seen in table 3, after eight iterations (step 8), the null hypothesis of equal coefficients is rejected for the variables backache, blood and other_disease. Hence, our final model consists of seven constrained and three varying variables.

Finally, as specification test, a global Wald test on the full model with constraints is applied confirming the null hypothesis that the parallel regression assumption is not violated (see table 4). In the example, the result of the autofit procedure with three varying and seven constrained variables meets the parallel-lines assumption. Thus, in contrast to the full varying model (see table 2), this specification is preferable and reflects best the observable heterogeneity in the data.

Table 4: Specification test

Wald test of parallel lines assumption for the final model:	
(1)	[mleq1]mental - [mleq2]mental = 0
(2)	[mleq1]gastric_ulcer - [mleq2]gastric_ulcer = 0
(3)	[mleq1]stroke - [mleq2]stroke = 0
(4)	[mleq1]cancer - [mleq2]cancer = 0
(5)	[mleq1]chol - [mleq2]chol = 0
(6)	[mleq1]heart - [mleq2]heart = 0
(7)	[mleq1]pul_asthma - [mleq2]pul_asthma = 0
(8)	[mleq1]mental - [mleq3]mental = 0
(9)	[mleq1]gastric_ulcer - [mleq3]gastric_ulcer = 0
(10)	[mleq1]stroke - [mleq3]stroke = 0
(11)	[mleq1]cancer - [mleq3]cancer = 0
(12)	[mleq1]chol - [mleq3]chol = 0
(13)	[mleq1]heart - [mleq3]heart = 0

(14) [mleq1]pul_asthma - [mleq3]pul_asthma = 0

(15) [mleq1]mental - [mleq4]mental = 0

(16) [mleq1]gastric_ulcer - [mleq4]gastric_ulcer = 0

(17) [mleq1]stroke - [mleq4]stroke = 0

(18) [mleq1]cancer - [mleq4]cancer = 0

(19) [mleq1]chol - [mleq4]chol = 0

(20) [mleq1]heart - [mleq4]heart = 0

(21) [mleq1]pul_asthma - [mleq4]pul_asthma = 0

chi2(21)	= 17.57
Prob > chi2	= 0.6758

Notes: Notes: An insignificant test statistic indicates that the final model does not violate the parallel lines assumption.

The final results of the procedure are displayed in table 5. Backache is highly significant throughout the categories. However, the negative effect is strongest for equation 3 (categories 1-3 vs. 4-5). Again, the variable blood shows only a significant impact for equations 3 and 4 and other_disease is highly significant for all categories. The main difference between a model with full variation and the preferred approach are the constrained variables. For instance, cancer now shows a general significant impact while in table 2, it only has a significant effect in equation 2. For other variables like chol, mental, pul_asthma and stroke, the difference is now that these variables are significantly negative for all categories. Hence, our findings suggest that the model with full variation is overspecified. The results produced with the autofit option show that for some variables, there exists significant variation throughout the reported categories. To sum up, the three variables blood, backache and other_disease drive the observed heterogeneity in our dependent variable self-assessed health.

Table 5: Regoprob2 with autofit

Random-Effects Generalised Ordered Probit				Number of obs	= 1186
				Wald chi2(19)	= 161.14
Log likelihood = -1157.435				Prob>chi2	= 0.0000
sah	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
mleq1					

backache	-0.9735	0.2911	-3.34	0.001	-1.54405	-0.40303
blood	0.2265	0.2832	0.80	0.424	-0.32853	0.78156
cancer	-0.6168	0.2555	-2.41	0.016	-1.11768	-0.11596
chol	-0.3526	0.1243	-2.84	0.005	-0.59626	-0.10899
gastric_ulcer	-0.4150	0.2729	-1.52	0.128	-0.94988	0.11997
heart	-0.8726	0.1620	-5.39	0.000	-1.19007	-0.55506
mental	-0.6034	0.1800	-3.35	0.001	-0.95615	-0.25063
other_disease	-1.0697	0.2902	-3.69	0.000	-1.63845	-0.50104
pul_asthma	-0.8423	0.1891	-4.45	0.000	-1.21295	-0.47156
stroke	-0.8008	0.2634	-3.04	0.002	-1.31700	-0.28467
_cons	4.2281	0.4217	10.03	0.000	3.40169	5.05456
mleq2						
backache	-0.6372	0.1640	-3.88	0.000	-0.95873	-0.31569
blood	-0.1302	0.1566	-0.83	0.406	-0.43718	0.17680
cancer	-0.6168	0.2555	-2.41	0.016	-1.11768	-0.11596
chol	-0.3526	0.1243	-2.84	0.005	-0.59626	-0.10899
gastric_ulcer	-0.4150	0.2729	-1.52	0.128	-0.94988	0.11997
heart	-0.8726	0.1620	-5.39	0.000	-1.19007	-0.55506
mental	-0.6034	0.1800	-3.35	0.001	-0.95615	-0.25063
other_disease	-0.9224	0.1586	-5.81	0.000	-1.23333	-0.61150
pul_asthma	-0.8423	0.1891	-4.45	0.000	-1.21295	-0.47156
stroke	-0.8008	0.2634	-3.04	0.002	-1.31700	-0.28467
_cons	2.7693	0.2336	11.85	0.000	2.31139	3.22725
mleq3						
backache	-1.3741	0.1643	-8.37	0.000	-1.69599	-1.05213
blood	-0.6849	0.1295	-5.29	0.000	-0.93861	-0.43111
cancer	-0.6168	0.2555	-2.41	0.016	-1.11768	-0.11596
chol	-0.3526	0.1243	-2.84	0.005	-0.59626	-0.10899
gastric_ulcer	-0.4150	0.2729	-1.52	0.128	-0.94988	0.11997
heart	-0.8726	0.1620	-5.39	0.000	-1.19007	-0.55506

mental	-0.6034	0.1800	-3.35	0.001	-0.95615	-0.25063
other_disease	-1.4019	0.1524	-9.20	0.000	-1.70062	-1.10317
pul_asthma	-0.8423	0.1891	-4.45	0.000	-1.21295	-0.47156
stroke	-0.8008	0.2634	-3.04	0.002	-1.31700	-0.28467
_cons	1.3283	0.1217	10.91	0.000	1.08970	1.56685
mleq4						
backache	-1.2852	0.3676	-3.50	0.000	-2.00569	-0.56473
blood	-0.4003	0.1949	-2.05	0.040	-0.78238	-0.01828
cancer	-0.6168	0.2555	-2.41	0.016	-1.11768	-0.11596
chol	-0.3526	0.1243	-2.84	0.005	-0.59626	-0.10899
gastric_ulcer	-0.4150	0.2729	-1.52	0.128	-0.94988	0.11997
heart	-0.8726	0.1620	-5.39	0.000	-1.19007	-0.55506
mental	-0.6034	0.1800	-3.35	0.001	-0.95615	-0.25063
other_disease	-0.8437	0.2422	-3.48	0.000	-1.31844	-0.36893
pul_asthma	-0.8423	0.1891	-4.45	0.000	-1.21295	-0.47156
stroke	-0.8008	0.2634	-3.04	0.002	-1.31700	-0.28467
_cons	-1.3403	0.1411	-9.50	0.000	-1.61689	-1.06365
rho						
_cons	0.4393	0.0835	5.26	0.000	0.27567	0.60290

4 CONCLUSIONS

In the empirical analysis of categorical dependent variables, the problems associated with the parallel-lines assumption should be taken into account. To deal with this, knowledge about the effects of the explanatory variables on the different categories is needed. An analysis based on an underlying theory, that provides information about the variables that violate the parallel-lines assumption would be preferable. But in most cases that is not the case. With the autofitting procedure implemented in regoprob2, we suggest a pragmatic and empirically robust approach to identify the variables that should be constrained. Furthermore, to the best knowledge of the authors, this is the first application of this kind for panel data. Taking into account that a standard ordered probit model may violate the parallel-lines

assumption and that a full-variation model is often overspecified, in absence of theory based advice an iterative procedure like autofit could be seen as the "lesser of three evils". In our example, we show in how far a variable such as self-assessed health is prone to observed heterogeneity. If one does not account for this, any varying effects of the explanatory variables on the categories will be neglected in the standard ordered probit model. Accordingly, our regoprob2 command combines the detection of observed heterogeneity in categorical variables with the inclusion of unobserved individual heterogeneity using a random-effects estimator.

ACKNOWLEDGEMENTS

Stefan Boes of the University of Zurich wrote regoprob and kindly gave permission to use parts of his code for regoprob2. See regoprob for a description of the former regoprob command.

Richard Williams of the Notre Dame Department of Sociology wrote gologit2 and kindly gave permission to use parts of his code for programming regoprob2. For a more detailed description of gologit2 and its features, see Williams (2006).

CITATION OF THE SOFTWARE MODULE

regoprob2 is not an official Stata command. It is a free contribution to the research community - like a paper - and available on SSC archive. Please cite it as:

Pfarr, C., Schmid, A. and U. Schneider (2010), REGOPROB2: Stata module to estimate random-effects generalized ordered probit models (update), Statistical Software Components, Boston College Department of Economics.

REFERENCES

Boes, S. 2007. Three Essays on the Econometric Analysis of Discrete Dependent Variables. Zurich: University of Zurich.

Boes, S. and R. Winkelmann. 2006. "Ordered Response Models." Allgemeines Statistisches Archiv 90: 167–181.

Börsch-Supan, A., M. Coppola, L. Essig, A.Eymann, and D. Schunk. 2008. "The German SAVE Study - Design and Results"mea studies 06. Mannheim Research Institute for the Economics of Aging, Mannheim.

Frechette, G. R. 2001. "sg158: Random-Effects Ordered Probit." Stata Technical Bulletin 59: 23–27.

Greene, W. H., M.N. Harris, B. Hollingsworth, and P.Maitra. 2008. "A Bivariate Latent Class Correlated Generalised Ordered Probit Model with an Application to Modeling Observed Obesity Levels", Working Paper Nr. 08-18.New York University, Department of Economics, New York.

Greene, W. H. and D.A. Hensher. 2010. Modeling ordered choices, A primer. Cambridge:Cambridge University Press.

Long, J. S. 1997. Regression models for categorical and limited dependent variables. Thousand Oaks, Calif: Sage Publ.

Pfarr, C., A. Schmid, and U. Schneider. 2010. "REGOPROB2: Stata module to estimate random-effects generalised ordered probit models" Statistical Software Components, Boston College Department of Economics.

Pudney, S. and M. Shields. 2000. "Gender, Race, Pay and Promotion in the British Nursing Profession, Estimation of a Generalised Ordered Probit Model." Journal of Applied Econometrics 15: 367–399.

Schneider, U., C. Pfarr, B. S. Schneider, and V. Ulrich. 2011. "I feel good. Gender differences and reporting heterogeneity in self-assessed health" European Journal of Health Economics, forthcoming.

Williams, R. 2006. "Generalised ordered logit/partial proportional odss models for ordinal dependent variables." Stata Journal 6: 58–82.

APPENDIX

Table A1: Variable description

Variable name	Label
health	self-assessed health, 1=very bad, 5=very good
backache	1, ifchronicbackache
blood	1, if individual suffer from hypertension
cancer	1, if individual is diagnosed with cancer
chol	1, if individual has a higher cholesterol level
gastric_ulcer	1, if a gastric ulcer is diagnosed
heart	1, if individual suffers heart diseases
mental	1, if mental disorders
other_disease	1, if other diseases
pul_asthma	1, if chronic chest disease or asthma
stroke	1, if circulatory disorders or stroke