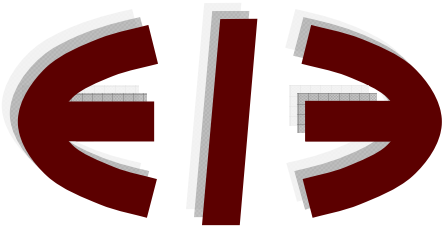
EERI
Economics and Econometrics Research Institute

Estimating ordered categorical variables using panel data: a generalized ordered probit model with an autofit procedure

Christian Pfarr, Andreas Schmid and Udo Schneider

EERI Research Paper Series No 43/2010

ISSN: 2031-4892



EERI
Economics and Econometrics Research Institute
Avenue de Beaulieu
1160 Brussels
Belgium

Tel: +322 299 3523
Fax: +322 299 3523
www.eeri.eu

Estimating ordered categorical variables using panel data: a generalized ordered probit model with an autofit procedure

Christian Pfarr – Andreas Schmid – Udo Schneider

June 2010

Correspondence address:
University of Bayreuth
Department of Law and Economics
Institute of Public Finance
D-95447 Bayreuth
Phone: +49-921-554324
Email: christian.pfarr@uni-bayreuth.de
andreas.schmid@uni-bayreuth.de
udo.schneider@uni-bayreuth.de

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 a random effects generalized ordered probit model. We combine the test procedure developed by Richard Williams (gologit2) with the random effects estimation command regoprob by Stefan Boes.

JEL: C23; C25; C87; I10

Keywords: generalized ordered probit; panel data; autofit, self-assessed health

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 such.

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.

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 constraint and meet the parallel-lines assumption. On the other hand, a fully flexible approach (generalized 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 generalized 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 Framework

When analyzing 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:

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

Here, κ_j are the unknown threshold parameters and β_s are the unknown coefficients.[1] The function F usually represents a cumulative 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:

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

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:

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

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, Harris, Hollingsworth 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

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

differ across the outcomes leads to a generalized ordered probit model. Here, the threshold parameters depend on the covariates:

$$\kappa_j = \tilde{\kappa}_j + x'\gamma_j, \tag{1.4}$$

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 generalized ordered probit model leads to the following expression:

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

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 generalized ordered probit model has one index $x'\beta_j$ for each category j of the outcome variable.[2] 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 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 generalized ordered probit approach (cf. Boes (2007), p. 133). More formally, let SAH be an ordinal variable which takes on the values j = 1,..., J. In contrast to the cross-section representation, the outcome probabilities are conditional on the individual effect α_i:[3]

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

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 generalized 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, differ-

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

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

ences 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 wanted. 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 generalized 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 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.[4] 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 generalized ordered probit model with the autofit option: an example

To discuss the estimation of a random effects generalized ordered probit model for ordered categorical variables we use self assessed health as dependent variable. 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.[5] For illustration purposes, we restrict the analysis to a 10 %-random sample of the original SAVE data[6] consisting of 1,186 individuals for the years 2006 to 2008.

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

[5] For more details regarding reporting heterogeneity in self-assessed health see Pfarr, Schneider, Schneider et al. (2010).

[6] 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 behavior focused

Table 1: Variable description

variable name	label
health	self assessed health, 1=very bad, 5=very good
backache	1, if chronic backache
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

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

Figure 1: Results of the fully constrained random effects ordered probit model.

Random Effects Ordered Probit — Number of obs = 1186
Log likelihood = -1176.8221 — LR chi2(10) = 415.84
Prob > chi2 = 0.0000

sah	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
eq1						
backache	-1.09901	.1295222	-8.49	0.000	-1.352869	-.8451511
blood	-.4476448	.1046306	-4.28	0.000	-.6527171	-.2425725
cancer	-.6490859	.2574948	-2.52	0.012	-1.153767	-.1444053
chol	-.3640996	.1257019	-2.90	0.004	-.6104708	-.1177284
gastric_ul~r	-.4358842	.2758477	-1.58	0.114	-.9765359	.1047674
heart	-.8273481	.1608416	-5.14	0.000	-1.142592	-.5121042
mental	-.5861793	.1808901	-3.24	0.001	-.9407174	-.2316412
other_dise~e	-1.217452	.1248299	-9.75	0.000	-1.462114	-.9727896
pul_asthma	-.859511	.1911378	-4.50	0.000	-1.234134	-.4848878
stroke	-.7893462	.2675926	-2.95	0.003	-1.313818	-.2648743
_cut1						
_cons	-4.703721	.3560371	-13.21	0.000	-5.401541	-4.005901
_cut2						
_cons	-3.280877	.2416687	-13.58	0.000	-3.754538	-2.807215
_cut3						
_cons	-1.159621	.1052621	-11.02	0.000	-1.365931	-.9533113
_cut4						
_cons	1.358256	.1316719	10.32	0.000	1.100184	1.616328
rho						
_cons	.4631909	.0775529	5.97	0.000	.3111901	.6151918

on savings and old-age provisions but also deals with aspects of health and health behavior (cf. Börsch-Supan, Coppola, Essig et al. (2008)).

In contrast to the results above, a generalized 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 figure 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 generalized 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 generalized 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.

Figure 2: Random effects generalized ordered probit with all variables varying.

Random Effects Generalized Ordered Probit — Number of obs = 1186
Log likelihood = -1145.8067 — LR chi2(40) = 22.08, Prob > chi2 = 0.9904

sah	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mleq1						
backache	-.9737055	.3137231	-3.10	0.002	-1.588591	-.3588195
blood	.0816322	.3132906	0.26	0.794	-.532406	.6956705
cancer	-.2651546	.6969769	-0.38	0.704	-1.631204	1.100895
chol	-.4152143	.3220881	-1.29	0.197	-1.046495	.2160668
gastric_ul~r	-.2362338	.8399071	-0.28	0.779	-1.882421	1.409954
heart	-.7720208	.3363741	-2.30	0.022	-1.431302	-.1127397
mental	-.8016768	.3648811	-2.20	0.028	-1.516831	-.0865229
other_dise~e	-1.154011	.3171672	-3.64	0.000	-1.775648	-.5323753
pul_asthma	-.9270324	.4121552	-2.25	0.024	-1.734842	-.1192229
stroke	-.2662729	.6011453	-0.44	0.658	-1.444496	.9119503
_cons	4.408939	.4940551	8.92	0.000	3.440609	5.377269
mleq2						
backache	-.6614166	.1734815	-3.81	0.000	-1.001434	-.3213992
blood	-.1545973	.1656071	-0.93	0.351	-.4791813	.1699867
cancer	-.9160748	.3508845	-2.61	0.009	-1.603796	-.2283538
chol	-.1535008	.1928042	-0.80	0.426	-.5313901	.2243886
gastric_ul~r	-.0508341	.4013878	-0.13	0.899	-.8375397	.7358715
heart	-.8606609	.21726	-3.96	0.000	-1.286483	-.4348391
mental	-.6307617	.2437958	-2.59	0.010	-1.108593	-.1529307
other_dise~e	-.9808445	.1695131	-5.79	0.000	-1.313084	-.648605
pul_asthma	-1.094238	.2604748	-4.20	0.000	-1.604759	-.5837164
stroke	-1.017224	.3392896	-3.00	0.003	-1.682219	-.3522283
_cons	2.876707	.2663424	10.80	0.000	2.354685	3.398728

mleq3						
backache	-1.429105	.1764258	-8.10	0.000	-1.774893	-1.083317
blood	-.677648	.1364458	-4.97	0.000	-.9450769	-.4102192
cancer	-.4145856	.3313946	-1.25	0.211	-1.064107	.2349359
chol	-.4047233	.1641518	-2.47	0.014	-.726455	-.0829917
gastric_ul~r	-.6336205	.4050083	-1.56	0.118	-1.427422	.1601812
heart	-1.148817	.2300788	-4.99	0.000	-1.599763	-.6978708
mental	-.5660157	.2466853	-2.29	0.022	-1.04951	-.0825213
other_dise~e	-1.455258	.1641961	-8.86	0.000	-1.777076	-1.133439
pul_asthma	-.7738617	.2395387	-3.23	0.001	-1.243349	-.3043744
stroke	-.7297693	.3658627	-1.99	0.046	-1.446847	-.0126915
_cons	1.380838	.1343809	10.28	0.000	1.117456	1.64422
mleq4						
backache	-1.516481	.4174031	-3.63	0.000	-2.334576	-.6983857
blood	-.4196904	.2087632	-2.01	0.044	-.8288587	-.0105221
cancer	-6.022424	387.4557	-0.02	0.988	-765.4217	753.3768
chol	-.7821387	.333523	-2.35	0.019	-1.435832	-.1284456
gastric_ul~r	-6.279077	430.5392	-0.01	0.988	-850.1205	837.5623
heart	-.4605701	.4015908	-1.15	0.251	-1.247674	.3265335
mental	-.7315414	.6360083	-1.15	0.250	-1.978095	.5150121
other_dise~e	-.8872761	.258128	-3.44	0.001	-1.393198	-.3813545
pul_asthma	.0785383	.4372637	0.18	0.857	-.7784827	.9355593
stroke	-5.746161	546.1385	-0.01	0.992	-1076.158	1064.666
_cons	-1.369723	.1539205	-8.90	0.000	-1.671401	-1.068044
rho						
_cons	.482422	.0845786	5.70	0.000	.316651	.648193

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.[7]

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 constraint. In step two, gastric_ulcer meets the parallel-lines assumption.

Figure 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)
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)
```

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

As can be seen in figure 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 that confirming the null hypothesis that the parallel regression assumption is not violated (see figure 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 figure 2), this specification is preferable and reflects best the observable heterogeneity in the data.

Figure 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

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 figure 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 figure 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 report-

ed categories. To sum up, the three variables blood, backache and other_disease drive the observed heterogeneity in our dependent variable self-assessed health.

Figure 5: Regoprob2 with autofit

Random Effects Generalized Ordered Probit — Number of obs = 1186
Log likelihood = -1157.4352 — Wald chi2(19) = 161.14
Prob > chi2 = 0.0000

sah	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mleq1						
backache	-.9735364	.291081	-3.34	0.001	-1.544045	-.4030281
blood	.2265152	.2831933	0.80	0.424	-.3285334	.7815638
cancer	-.6168198	.2555456	-2.41	0.016	-1.11768	-.1159597
chol	-.3526229	.1243047	-2.84	0.005	-.5962557	-.1089902
gastric_ul~r	-.4149588	.272926	-1.52	0.128	-.949884	.1199663
heart	-.8725646	.1619958	-5.39	0.000	-1.190071	-.5550586
mental	-.6033877	.1799832	-3.35	0.001	-.9561484	-.2506271
other_dise~e	-1.069742	.2901612	-3.69	0.000	-1.638447	-.5010366
pul_asthma	-.8422519	.1891343	-4.45	0.000	-1.212948	-.4715554
stroke	-.8008353	.2633551	-3.04	0.002	-1.317002	-.2846687
_cons	4.22812	.4216583	10.03	0.000	3.401685	5.054555
mleq2						
backache	-.6372132	.1640445	-3.88	0.000	-.9587345	-.3156919
blood	-.1301879	.1566306	-0.83	0.406	-.4371783	.1768025
cancer	-.6168198	.2555456	-2.41	0.016	-1.11768	-.1159597
chol	-.3526229	.1243047	-2.84	0.005	-.5962557	-.1089902
gastric_ul~r	-.4149588	.272926	-1.52	0.128	-.949884	.1199663
heart	-.8725646	.1619958	-5.39	0.000	-1.190071	-.5550586
mental	-.6033877	.1799832	-3.35	0.001	-.9561484	-.2506271
other_dise~e	-.9224133	.1586319	-5.81	0.000	-1.233326	-.6115005
pul_asthma	-.8422519	.1891343	-4.45	0.000	-1.212948	-.4715554
stroke	-.8008353	.2633551	-3.04	0.002	-1.317002	-.2846687
_cons	2.76932	.2336415	11.85	0.000	2.311391	3.227249
mleq3						
backache	-1.374059	.1642543	-8.37	0.000	-1.695992	-1.052127
blood	-.6848629	.1294674	-5.29	0.000	-.9386143	-.4311116
cancer	-.6168198	.2555456	-2.41	0.016	-1.11768	-.1159597
chol	-.3526229	.1243047	-2.84	0.005	-.5962557	-.1089902
gastric_ul~r	-.4149588	.272926	-1.52	0.128	-.949884	.1199663
heart	-.8725646	.1619958	-5.39	0.000	-1.190071	-.5550586
mental	-.6033877	.1799832	-3.35	0.001	-.9561484	-.2506271
other_dise~e	-1.401897	.1524138	-9.20	0.000	-1.700623	-1.103171
pul_asthma	-.8422519	.1891343	-4.45	0.000	-1.212948	-.4715554
stroke	-.8008353	.2633551	-3.04	0.002	-1.317002	-.2846687
_cons	1.328276	.1217234	10.91	0.000	1.089702	1.566849
mleq4						
backache	-1.285208	.3675969	-3.50	0.000	-2.005685	-.5647317
blood	-.4003325	.1949268	-2.05	0.040	-.7823821	-.0182829
cancer	-.6168198	.2555456	-2.41	0.016	-1.11768	-.1159597
chol	-.3526229	.1243047	-2.84	0.005	-.5962557	-.1089902
gastric_ul~r	-.4149588	.272926	-1.52	0.128	-.949884	.1199663
heart	-.8725646	.1619958	-5.39	0.000	-1.190071	-.5550586
mental	-.6033877	.1799832	-3.35	0.001	-.9561484	-.2506271
other_dise~e	-.8436839	.2422246	-3.48	0.000	-1.318435	-.3689325
pul_asthma	-.8422519	.1891343	-4.45	0.000	-1.212948	-.4715554
stroke	-.8008353	.2633551	-3.04	0.002	-1.317002	-.2846687
_cons	-1.340268	.1411368	-9.50	0.000	-1.616891	-1.063645
rho						
_cons	.4392843	.0834766	5.26	0.000	.2756732	.6028955

4 Conclusion

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.

5 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 goprobit. For a more detailed description of gologit2 and its features, see the reference below or gologit2.

6 References

Boes, S. (2007), *Three Essays on the Econometric Analysis of Discrete Dependent Variables,* Universität Zürich, Zürich.

Boes, S. and Winkelmann, R. (2006), Ordered Response Models, in: *Allgemeines Statistisches Archiv*, 90, pp. 167–181.

Börsch-Supan, A., Coppola, M., Essig, L., Eymann, A. and Schunk, D. (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, in: *Stata Technical Bulletin*, 59, pp. 23–27.

Greene, W. H., Harris, M. N., Hollingsworth, B. and Maitra, P. (2008), A Bivariate Latent Class Correlated Generalized 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 Hensher, D. A. (2010), *Modeling ordered choices, A primer,* Cambridge University Press, Cambridge.

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

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

Pfarr, C., Schneider, B. S., Schneider, U. and Ulrich, V. (2010), Self-assessed health, gender differences and reporting heterogeneity: empirical evidence using multiple imputed data, *Discussion Paper,* Nr. 03-10, University of Bayreuth, Department of Law and Economics, Bayreuth.

Pudney, S. and Shields, M. (2000), Gender, Race, Pay and Promotion in the British Nursing Profession, Estimation of a Generalized Ordered Probit Model, in: *Journal of Applied Econometrics*, 15(4), pp. 367–399.

Williams, R. (2006), Generalized ordered logit/partial proportional odss models for ordinal dependent variables, in: *Stata Journal*, 6(1), pp. 58–82.