Indirect Estimation of Linear Models with Ordinal Regressors. A Monte Carlo Study and some Empirical Illustrations.

Martin Kukuk[*]
University of Tübingen [†]

Abstract

This paper investigates the effects of ordinal regressors in linear regression models. Each ordered categorical variable is interpreted as a rough measurement of an underlying continuous variable as it is often done in microeconometrics for the dependent variable. It is shown that using ordinal indicators only leads to correct answers in a few special cases. In most situations, the usual estimators are biased. In order to estimate the parameters of the model consistently, the indirect estimation procedure suggested by Gourieroux et al. (1993) is applied. To demonstrate this method, first a simulation study is performed and then in a second step, two real data sets are used. In the latter case, continuous regressors are transformed into categorical variables to study the behavior of the estimation procedure. In general, the indirect estimators lead to adequate results.

KEY WORDS: Microeconometrics, Exogenous Variables with Ordinal Scale, Latent Variables, Indirect Estimation.

JEL-CLASSIFICATION: C2, C4.

[*] The data is supplied by the Centre for European Economic Research (ZEW, Mannheim) which is gratefully acknowledged. I thank Roman Liesenfeld, Winfried Pohlmeier, Gerd Ronning, Rüdiger Wapler, and two anonymous referees for helpful comments on an earlier draft.

[†] Address for correspondence: Wirtschaftswissenschaftliche Fakultät, Universität Tübingen, Mohlstr. 36, 72074 Tübingen, Germany. E-mail: Martin.Kukuk@uni-tuebingen.de.

1 Introduction

In sample surveys on the individual level (e.g. households or firms), it is often the case that many questions are asked categorically. This is due to the fact that it is less time consuming to answer, for instance, whether ones annual income falls into a specific income class rather than giving the exact value. Microeconometric models are available in cases where the dependent variable carries limited information. However, when explaining such a variable using other variables from the same survey, it is likely that those explanatory variables also carry only limited information.

As an example, Li (1977) wants to explain the individual propensity of homeownership as a linear function of household income, age of head, family size, and race of head. In essence, he applies a binary logit model, since observations on the dependent variable are only available as homeownership status. The explanatory variables *income* and *age* are also measured categorically in this survey. Income is measured using 4 categories. Li (1977) uses a set of three dummy variables to measure the influence of income on homeownership. The same applies to the age variable where a set of 4 dummy variables is included to represent the 5 age categories. This has become common practice (e.g. Theil, 1971 p. 633ff., McIntosh et al., 1989 p. 255). In the latter paper it is mentioned, however, that this common practice treats ordered variables on the left–hand side and right hand–side asymmetrically.

Throughout this paper we apply K. Pearson' s (1901) idea of an underlying continuous variable for an ordinal indicator to also explanatory variables. This idea, applied to the left–hand side variable, is the basis, for instance, in the ordered probit/logit model. Given this assumption, we will show that for models which are formulated linearly in the continuous variables, the common practice of replacing an ordinal indicator by a set of dummy variables or using the ordinal indicator itself as a regressor could lead to wrong answers with respect to whether or not a continuous latent variable has a significant influence.

Our approach is conceptually different from non-linear regression models with discrete explanatory variables (e.g. Bierens and Hartog, 1988). Those models are formulated conditional on the observed ordered variables. Therefore, the results have to be interpreted in terms of this measurement level. This implies, for instance, for the above mentioned example that not whether *income* has an effect on homeownership is tested, but instead, whether the predefined income classes effect the outcome.

The paper is organized as follows: In Section 2, the problem is discussed in more detail. Using a linear regression model the effect of regressor variables with only limited information is demonstrated. In Section 3, the indirect estimation procedure is introduced as a possibility of estimating the parameters of the latent model. With this method, the latent model is first simulated depending on the parameter of interest. Then the loss of information due to categorizing the continuous variables is imitated in order to have the same kind of observations as in the data set at hand. An auxiliary model is estimated using both data sets. The parameters of interest are calibrated in order to obtain close auxiliary parameter estimates. A simulation study is performed in Section 4 to compare the indirect estimation method with the Ordinary Least Squares (OLS) estimator in which the categorical information is used directly. In Section 5, the comparison between those two methods is continued by using real data sets. Within this experiment, some regressors are categorized to have a "true" benchmark. The dependent variable is also categorized to demonstrate the usefulness of the method in such a setting. Finally, in section 6 some conclusions are made and further applications are discussed.

2 Modeling Ordinal Regressors

Starting point for the discussion in this paper is the linear model

$$y^* = \beta_0 + x_1^*\beta_1 + x_2^*\beta_2 + \varepsilon \quad . \tag{1}$$

Microeconometrics as a special field in econometrics evolved due to the fact that the dependent variable y^* cannot always be observed directly. For instance, if it is only known that y^* is greater or less than a fixed value c, meaning that we only observe

$$y = \begin{cases} 1 & \text{if } y^* \geq c \\ 0 & \text{otherwise} \end{cases}$$

we can apply the *binary probit* or *binary logit* model. Another example is the measurement equation[1]

$$y = k \quad \Longleftrightarrow \quad \alpha_{k-1} \leq y^* < \alpha_k \qquad k = 1, 2, \ldots, K \quad . \tag{2}$$

leading to the *ordered probit/logit* model (Ronning, 1991 Chap. 2). The latter two models differ in the distributional assumption on ε and hence the conditional distribution of $y^*|\boldsymbol{x}^*$.

[1] Throughout the paper the superscript $*$ is used to indicate a continuous variable, whereas for its ordered counterpart this superscript is omitted.

Due to the loss of information in the observed dependent variable, not all of the model parameters are identified. Restricting $\beta_0 = 0$ and $\sigma_\varepsilon = 1$ identifies the model. This can be interpreted as estimating the thresholds $\alpha_1, \ldots, \alpha_{k-1}$ of equation (2) and the parameters β_i' of

$$\frac{y^*}{\sigma_\varepsilon} = x_1^*\beta_1' + x_2^*\beta_2' + \varepsilon' \tag{3}$$

with $\beta_i' = \beta_i/\sigma_\varepsilon$ instead of equation (1). It should be mentioned that this restriction makes it impossible to test a linear restriction in terms of the original parameters, i. e. $\beta_i = k$; however, testing $\beta_i = 0$, which is our main concern in this paper, is not affected by the restricting assumption.

Next, assume the right hand variable x_2^* in equation (1) is not observed directly whereas y^* and x_1^* are continuously measurable[2]. Variable x_2^* is assumed to be measured qualitatively as x_2 analogous to equation (2). Hsiao and Mountain (1985), Ross (1987), and Ross and Zimmermann (1993) suggest assuming a normal distribution[3] for the latent variable x_2^* and derive the conditional expected value $\mathrm{E}(x_2^* \mid x_2)$ which can be interpreted as a general residual (Gourieroux et al., 1987 p.14.):

$$\mathrm{E}(x_2^* \mid x_2 = j) = \frac{\phi(\alpha_{j-1}) - \phi(\alpha_j)}{\Phi(\alpha_j) - \Phi(\alpha_{j-1})} \quad , \tag{4}$$

where $\phi(\cdot)$ denotes the density and $\Phi(\cdot)$ the distribution function of the standard normal distribution. The unobserved residual depending on the realization of x_2^* is then[4]

$$\xi_2 = x_2^* - \mathrm{E}(x_2^* \mid x_2) \quad . \tag{5}$$

The conditional distribution function of $x_2^*|x_2 = j$ is a truncated normal distribution with support $[\alpha_{j-1}; \alpha_j[$. Inserting (5) into equation (1) yields the well known errors–in–variables (EIV) problem leading to biased and inconsistent OLS–estimates. Additionally, heteroscedasticity occurs (Yatchew and Griliches, 1984) since the variance $\mathrm{var}(x_2^* \mid x_2)$ depends on x_2. Therefore, Hsiao and Mountain (1985) suggest using $(x_1^*; \mathrm{E}(x_2^*|x_2))$ as instruments for $(x_1^*; x_2^*)$ and obtain consistent IV–estimators. However, some problems arise due to unknown covariance parameters which can only be solved using additional assumptions. Kao and Schnell (1987) assume that these covariances are known.

For the special case of $\beta_2 = 0$ the use of $\mathrm{E}(x_2^*|x_2)$ instead of x_2^* leads to unbiased OLS–estimates and under this hypothesis the usual tests are applicable. In this situation, the regressor variables

[2] See also Ronning and Kukuk (1996).

[3] To be more precise, a standard normal distribution is assumed, implying that instead of β_2 in (1) one estimates $\beta_2'' = \beta_2 \cdot \sigma_{x_2^*}$. Additionally, the constant term changes to $\beta_0 - \beta_2''\mu_{x_2^*}$.

[4] It should be noted that equation (5) can also be formulated using the coding scheme $\tilde{\xi}_2 = x_2^* - x_2$.

are not correlated with the error term; hence, the significance level of a test $H_0 : \beta_2 = 0$ is correct. This is of great importance for practical purposes since the categorical indicator x_2 can be used to test the significance of β_2 in the latent model (1).

Otherwise, if for both right hand side variables x_1^* and x_2^* only categorical observations are available analogous to (2), equation (1) can be written as

$$y^* = \mathrm{E}(x_1^*|x_1)\beta_1 + \mathrm{E}(x_2^*|x_2)\beta_2 + \varepsilon + \xi_1\beta_1 + \xi_2\beta_2 \quad , \tag{6}$$

where the new error term $\nu = \varepsilon + \xi_1\beta_1 + \xi_2\beta_2$ will be correlated with the regressors in most cases. The distribution of the error term ν is a mixture of a normal and two truncated normal distributions which aggravates the use of standard EIV models. Even for the above mentioned case of $\beta_2 = 0$, the OLS–method using the observations $(y^*, \mathrm{E}(x_1^*|x_1), \mathrm{E}(x_2^*|x_2))$ lead to biased estimates and incorrect significance levels. Nevertheless, Nerlove et al. (1993) and Ross and Zimmermann (1993) apply this method by arguing that for low correlations between the regressors, the biases will be small. Simulation studies are mentioned to support their view. However, we will show in section 4 that the bias in the significance level is not of negligible order.

The usual way of estimating the parameter vector $\boldsymbol{\beta}$ correctly is to assume a multivariate distribution for y^*, x_1^*, x_2^*. For instance, if we assume a trivariate (standard–)normal distribution the ML–estimates for $\boldsymbol{\beta}$ can be derived from the ML–estimates of the covariance parameters of the normal distribution[5]. This is due to the fact that the regression parameters are a function of the covariance parameters of the joint distribution. Within this procedure it is also possible to only have qualitative observations y instead of y^*. However, problems arise if dummy variables should be included on the right side of equation (1). It is hard to imagine a mixed continuous–discrete joint distribution for all the variables involved and then derive a linear regression from it. A possible estimation strategy could be a reformulation of equation (1) into a system of linear equations and estimating it using e.g. MECOSA (Schepers and Arminger, 1992). However, this strategy requires slightly different distributional assumptions.

3 Indirect Estimation

In this paper we want to follow another procedure to estimate the parameters of equation (1) which allows the use of nominal scaled dummy variables on the right hand side. This method is based on the idea of *indirect estimation* proposed by Gourieroux et al. (1993) and Gallant and

[5] See Browne and Arminger (1994), Jöreskog (1990), or Kukuk (1991)

Tauchen (1996). This procedure uses simulation techniques which have become more and more attractive for practical purposes due to increases in computing power in recent years.

In the last section it was mentioned that, since x_1^* and x_2^* are not observable, a distributional assumption is necessary to derive conditional moments which are essential for the estimation of the model. For instance, if we assume that x_1^* and x_2^* follow a bivariate normal distribution, then realizations of these variables can be simulated. Together with simulated realizations for the residual ε and given values for the β_i's, simulated values for y^* are determined. The simulated values for x_i^* can be transformed into values of x_i for some given values of α_j according to the measurement equation (2). At this point we have simulated observations for (y^*, x_1, x_2) which of course depend on model parameters such as β_i and α_j. The measurement levels of these simulated data correspond to the realized observations from the survey. If the model is true and all the parameters are known, then realized observations at hand (y^*, x_1, x_2) follow the same distribution as the simulated data.

The basic idea of the indirect estimation method is to use an *auxiliary model.* In our context, such an auxiliary model could be a regression of y^* on the categorical indicators x_1, x_2:

$$y^* = \theta_0 + x_1\theta_1 + x_2\theta_2 + \eta \quad .$$

The resulting OLS–estimator $\hat{\boldsymbol{\theta}}$ is a biased estimator for the parameter vector $\boldsymbol{\beta}$ as shown above. The auxiliary parameter vector $\boldsymbol{\theta}$, which, in our context, has no meaningful interpretation, is estimated using the real data (denoted by $\hat{\boldsymbol{\theta}}$) and also using the simulated data (denoted by $\tilde{\boldsymbol{\theta}}$). Again, if the latent model is true and all the parameters known, the distribution of the $\hat{\boldsymbol{\theta}}$ is the same as the distribution of $\tilde{\boldsymbol{\theta}}$.

However, the model parameters are usually unknown. The estimates $\tilde{\boldsymbol{\theta}}$ from the simulated data[6] are a function of the model parameters. Therefore, the next step of the indirect estimation consists of a calibration of model parameters so that $\tilde{\boldsymbol{\theta}}$ is close to $\hat{\boldsymbol{\theta}}$. This is done by a Minimum–Distance step which in our model could be

$$\min_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\theta}}(\boldsymbol{\beta}) - \hat{\boldsymbol{\theta}})'(\tilde{\boldsymbol{\theta}}(\boldsymbol{\beta}) - \hat{\boldsymbol{\theta}}) \quad .$$

In our case, this minimization is performed using the GAUSS application module OPTMUM.

Gourieroux et al. (1993) show the consistency and asymptotic normality of the indirect estimator. In a couple of simulation studies they show for models in which exact ML-estimators are available

[6] For a given set of model parameters, more than one simulated data set can be simulated and for each of them we can estimate the auxiliary model. As a result, $\tilde{\boldsymbol{\theta}}$ could be taken as the mean of these estimates.

that the indirect estimator performs just as well as the ML-procedures in terms of standard deviations and root–mean–squared–errors.

In the next section we perform a simulation study to show the advantages of the indirect estimation over the crude way of using $\hat{\boldsymbol{\theta}}$ as an estimator for $\boldsymbol{\beta}$ which is often encountered in practical applications.

4 Simulation study

For our example, equation (1) is extended by two continuous and one dummy variable. The latter is constructed to have a positive correlation with the other explanatory variables. Now, the model is:

$$y^* = \beta_0 + x_1^*\beta_1 + x_2^*\beta_2 + x_3^*\beta_3 + x_4^*\beta_4 + D\beta_5 + \varepsilon \quad , \tag{7}$$

where the variables x_3^* and x_4^* will be treated as unobservable variables for which only categorical information is available. D is a dummy variable coded as 0 and 1. The vector of the continuous variables x_i^* is assumed to follow a multivariate standard normal distribution[7] with zero expectation and correlation matrix R:

$$R = \begin{pmatrix} 1 & .4 & .5 & .3 \\ & 1 & .4 & .35 \\ & & 1 & .4 \\ & & & 1 \end{pmatrix}$$

The values of the correlations are chosen to have magnitudes which can be observed in real data situations. Setting the parameter vector $\boldsymbol{\beta} = (\beta_0, \dots, \beta_5)' = (0, .4, .4, 0, .4, .3)'$ and using normally distributed residuals we obtain y^*. It should be noted that x_3^*, for which later on only the categorical information will be used, has no effect on y^* in this setting[8]. The variables x_3^* and x_4^* are categorized analogous to equation (2) in x_3 and x_4, respectively, each having three categories. Different values for the thresholds are used. The setting is as follows: we perform 1.000 simulation runs. In each run 1.000 observations[9] for all right hand variables of model (7) are simulated. For each data set 3 simulated data sets are used for the indirect estimation.

[7] This assumption is replaced at the end of this section by a multivariate t-distribution to assess the robustness of the estimation procedure.

[8] However, the correlation between y^* and x_3^* is approximately 0.53.

[9] Simulations with 500 observations were run as well leading to similar results.

In table 1 arithmetic means and standard deviations of the estimated parameters are recorded for different sets of thresholds. Results for three common practice procedures applying OLS are recorded as well as those for the indirect estimation procedure. OLS1 denotes a regression of y^* on x_1^*, x_2^*, x_3, x_4, and D; OLS2 is the same except x_3 and x_4 are replaced by $\mathrm{E}(x_3^*|x_3)$ and $\mathrm{E}(x_4^*|x_4)$, respectively; OLS3 is the procedure described in the first section using a set of two dummies for each ordered indicator each having 3 categories. In the latter case significance of x_3^* is usually tested by an F-test for both regression parameters. The results for all the OLS methods show more or less biased estimates for the parameter vector $\boldsymbol{\beta}$ which is due to the correlations between the regressors and the error term as shown above. The size of those biases depends on the values of the correlation between the regressors which are chosen as not too large in this example. The biases also depend on the threshold values which can be drastically seen in the last threshold setting.

Since the OLS–estimates for β_3 show biases it is not surprising that the empirical significance level is well above the true value of 5%. This means that by using the categorical indicators or a set of dummy variables, the null hypothesis that x_3^* has no influence on the dependent variable is rejected too often. In other words, it is stated too often that variable x_3^* has a significant influence on y^*.

No systematic bias can be observed for the indirect estimation method. On average, the true structure of the model is obtained so that the extra computational effort seems to be justified. The standard deviations of the parameters are larger than those resulting from OLS–estimation. However, this does not support the use of OLS since it produces large biases. To obtain a valid benchmark, those results should be compared to full information ML–estimates which are not available for our general model allowing the inclusion of dummy variables (Hsiao and Mountain, 1985). However, if we modify our model under consideration (7) by dropping the dummy variable D, leading to

$$y^* = \beta_0 + x_1^*\beta_1 + x_2^*\beta_2 + x_3^*\beta_3 + x_4^*\beta_4 + \varepsilon \quad , \tag{7'}$$

a Full–Information ML estimator is available (Kukuk, 1991). The same Monte–Carlo design as above is used for this modified model with the exception that we only consider the first set of thresholds. Additionally, the number of latent data sets used in the indirect estimation procedure is varied to demonstrate its effect on efficiency.

The results in table 2 show that the number of simulated data sets used in each indirect estimation has a considerable effect. The relative efficiency suggested by Krämer (1980), which is defined

Table 1: Simulation results for $\boldsymbol{x}^*$ following a multivariate normal distribution $N(0, R)$.

		Method	β_0	β_1	β_2	β_3	β_4	β_5	Reject $\beta_3 = 0$
Thresholds		True Values	0	0.4	0.4	0	0.4	0.3	in %
-1	1.2	OLS1	-.529	.412	.421	.037	.509	.310	19.1
-0.7	0.8		(.041)	(.017)	(.017)	(.034)	(.024)	(.086)	
		OLS2	.000	.412	.422	.023	.381	.310	18.8
			(.018)	(.017)	(.017)	(.021)	(.019)	(.086)	
		OLS3							15.7
		Ind. Est.	-.001	.401	.400	-.000	.399	.314	5.1
			(.019)	(.031)	(.028)	(.033)	(.024)	(.134)	
-0.9	0.4	OLS1	-.467	.410	.420	.026	.451	.313	17.6
-0.5	0.6		(.037)	(.017)	(.017)	(.025)	(.023)	(.056)	
		OLS2	-.001	.410	.421	.020	.380	.318	17.6
			(.019)	(.017)	(.017)	(.020)	(.020)	(.056)	
		OLS3							13.7
		Ind. Est.	-.001	.399	.400	.001	.398	.317	4.5
			(.021)	(.030)	(.028)	(.028)	(.026)	(.087)	
-0.5	0.5	OLS1	-.456	.411	.420	.022	.451	.315	16.5
-0.5	0.6		(.032)	(.017)	(.017)	(.023)	(.022)	(.058)	
		OLS2	-.001	.410	.421	.019	.380	.317	16.4
			(.019)	(.017)	(.017)	(.020)	(.020)	(.058)	
		OLS3							11.4
		Ind. Est.	-.001	.400	.399	-.001	.398	.318	6.0
			(.020)	(.030)	(.028)	(.027)	(.025)	(.091)	
-1.6	1.9	OLS1	-.769	.422	.436	.072	.612	.272	21.2
-1.5	0.8		(.067)	(.017)	(.017)	(.059)	(.029)	(.157)	
		OLS2.	.001	.423	.436	.034	.376	.274	20.5
			(.018)	(.017)	(.017)	(.028)	(.019)	(.158)	
		OLS3							15.7
		Ind. Est.	-.001	.399	.398	.004	.400	.275	4.3
			(.018)	(.035)	(.032)	(.053)	(.028)	(.244)	
-1.5	0.8	OLS1	-.855	.433	.462	.108	.752	.271	82.9
-1.6	1.9		(.068)	(.019)	(.019)	(.038)	(.056)	(.150)	
		OLS2	.002	.433	.462	.067	.356	.228	83.0
			(.018)	(.019)	(.019)	(.023)	(.028)	(.151)	
		OLS3							76.5
		Ind. Est.	-.001	.399	.398	.004	.400	.275	4.3
			(.018)	(.035)	(.032)	(.053)	(.028)	(.244)	

Table 2: Estimated Standard Errors of FIML and Indirect Estimation of model (7′)

Method	S.E. of Parameter				Efficiency
	β_1	β_2	β_3	β_4	
ML	.0188	.0166	.0212	.0167	
Ind3	.0258	.0266	.0288	.0215	.510
Ind6	.0238	.0209	.0250	.0203	.665
Ind12	.0213	.0190	.0228	.0183	.812
Ind18	.0206	.0192	.0218	.0188	.837
Ind24	.0196	.0190	.0219	.0186	.863

Note: Ind3 stands for Indirect Estimation using 3 simulated data sets. Efficiency is defined in (8).

by

$$\text{eff.} = \frac{\operatorname{tr}\left(\operatorname{Cov}(\hat{\theta}_{ML})\right)}{\operatorname{tr}\left(\operatorname{Cov}(\hat{\theta}_{Ind.Est.})\right)} \quad , \tag{8}$$

increases form 51%, using 3 simulated data sets, to 86.3% using 24 simulated data sets. In all situations, the means of the estimated parameters do not show any systematic deviation from their true values. Therefore, they are not reported in the table. The reported standard errors for the indirect estimation procedure in table 2 as well as table 1 are obtained by using the same random numbers in the indirect estimation procedure in all simulation runs. If we used a new set of random numbers in the indirect estimation for each simulation run, additional variation would come into play resulting in slightly higher standard errors of the parameters.

In order to simulate the latent model (7) with the indirect estimation procedure, the correlation structure between the continuous variables x_i^* must be estimated. In our procedure we assume a multivariate normal distribution to estimate the polychoric correlation (Olsson, 1979, and Kukuk, 1991 and 1994) between two latent variables and an estimator suggested by Brillinger (1982) for the correlation between a continuous and a latent variable. This corresponds to the design of our simulation study. In a next step, model (7) is simulated using a multivariate t-distribution (Fang et al., 1990 p. 85ff., and Lee and Lam, 1988) for the continuous variables x_i^*. The correlation structure, the model parameters, and the distribution for ε are kept unchanged. The assumptions in the indirect estimation procedure now deviate from the true model in which the latent variables have an excess kurtosis of 0.96, 2.7, and 29.4 for the parameters m=10, m=6, and m=3, respectively. The results of this simulation setup are given in table 3. They still indicate that the indirect estimation procedure yields satisfying answers. First attempts to extend this robustness result to other members of the class of *elliptically symmetric distributions* are promising. These distributions can be found quite often in other circumstances (Stoker, 1986,

Table 3: Simulation results for $\boldsymbol{x}^*$ following a multivariate t-distribution $Mt(m, 0, R)$

		Method	β_0	β_1	β_2	β_3	β_4	β_5	Reject $\beta_3 = 0$
Thresholds		True Values	0	0.4	0.4	0	0.4	0.3	in %
-1	1.2	OLS1	-.536	.413	.426	.040	.511	.330	16.1
-0.7	0.8		(.047)	(.019)	(.020)	(.040)	(.028)	(.102)	
		Ind. Est.	-.002	.399	.403	-.001	.391	.329	3.6
m=	10		(.021)	(.031)	(.031)	(.035)	(.025)	(.149)	
		OLS1	-.536	.415	.429	.037	.514	.364	13.4
			(.050)	(.022)	(.021)	(.043)	(.030)	(.120)	
		Ind. Est.	-.002	.404	.406	-.004	.382	.360	3.9
m=	6		(.021)	(.032)	(.031)	(.038)	(.026)	(.158)	
		OLS1	-.561	.416	.444	.049	.521	.440	26.0
			(.083)	(.040)	(.043)	(.076)	(.038)	(.158)	
		Ind. Est.	-.004	.397	.409	.0010	.370	.446	4.0
m=	3		(.021)	(.049)	(.041)	(.058)	(.032)	(.196)	

and Ruud, 1986).

5 Experiments with two Real Data Sets

The simulation study obtained satisfying results for the indirect estimation procedure even in those situations where the distributional assumption in the estimation procedure deviates from the true distribution. To study the performance of this method more carefully we apply it to two real data sets. The aim is to show how robust it is if some of the required assumptions are violated.

5.1 Analysing Real Estate Values

In the textbook of Berenson and Levine (1992), data of a real estate survey of 322 homes in two suburban New York counties in 1990 is given. Besides the value of the estates, other characteristics are surveyed. We consider a regression of the real estate value (Y^* in 1000 US\$) on the variables *annual taxes* X_1^*, *number of bathrooms* X_2^*, *age of the house* X_3^*, and *lot size* X_4^*. The following OLS-estimates are obtained:

$$\begin{array}{lllllll} \hat{Y}^* \; = & 135.1 & +0.004 \cdot X_1^* & +30.64 \cdot X_2^* & -0.2413 \cdot X_3^* & -1.41 \cdot X_4^* & R^2 = 0.298 \\ & (13.4) & (1.89) & (9.1) & (-1.28) & (-4.18) & \end{array}$$

The variables X_1^* to X_4^* are all continuously measured although the variable *number of bathrooms* (X_2^*) only takes on the values 1, 1.5, ... 3.5. Nevertheless, this variable is considered to be continuous but the assumed normal distribution is obviously violated. Also variable *lot size* does not follow a normal distribution; it shows a left–skewed distribution with an excess kurtosis of 6.3. The following experiment is conducted: the variables *age of the house* and *lot size* are categorized into binary variables X_3 and X_4, respectively. For this we use different values for the thresholds starting at the 10% percentile of the according distributions and successively increase them until the 90% percentiles are reached. For each data set, the OLS-estimates of the regression of Y^* on X_1^*, X_2^*, X_3, and X_4 as well as the indirect estimators are calculated and the significance of X_3^* and X_4^* are recorded on the 5% level. For the OLS method this means that the significance of X_3 and X_4 is taken as a proxy for the corresponding latent variable. It should be mentioned that the correlation between the continuous observations on *age of the house* and *lot size* is -0.21, which is not very large.

In table 4 the relative frequencies are shown where the "true" influence is denoted by "0 –" meaning that variable X_3^* has no effect and X_4^* has a negative effect on the dependent variable. Even in this case where the distributional assumptions are obviously violated, the indirect estimation outperforms the usual practice using the OLS–method. If the parameter space is reduced more, the dominance of the indirect procedure gets even stronger. In this restricted analysis, the correlation for instance between the variables X_3 and X_3^* becomes sufficiently large and, consequently, the indirect method almost always yields the "true" constellation 0– whereas OLS only obtains this constellation in 35% of the cases and in all other cases estimates significant negative effects of both variables X_3 and X_4. From this example we can infer that the indirect procedure is quite robust against violations of the distributional assumptions. Another result is, that the indirect estimation method is more precise the higher the correlation between the latent variable and the corresponding categorical indicator which solely depends on the threshold parameters. Of course, in practical applications these parameters cannot be chosen. Unless some prior information exists, as would be the case for the income distribution. By designing a survey questionnaire, this prior information can be used to determine income classes so that the resulting distribution of the qualitative variable has the desired shape.

5.2 Innovation Activities

As a second example a data set on innovations in the service sector is used which was conducted by the Centre for European Economic Research together with Fraunhofer–Institute ISI and infas Sozialforschung in 1995 (Licht et al., 1997). In this survey, approximately 3.000 enterprises

Table 4: Relative frequencies of significant parameters β_3 and β_4 in %

		OLS		
		β_3 β_4	β_3 β_4	
		0 –	– –	
Indirect	0 –	22.9	30.5	53.4
Estim.	– –	2.6	8.2	10.8
	0 0	9.2	12.4	27.6
		34.7	55.8	

Note: Marginal frequencies differ due to some minor cases not shown.

participated. Most questions in the questionnaire were designed to obtain qualitative answers since objective measuring scales are lacking. The empirical example to be shown in this section just serves demonstrative purposes and should not be interpreted as a meaningful specification. First, we regress the continuous variable *turnover* (standardized) on *number of employees*, which for computational reasons is also standardized, the categorical variable *size class* and the binary indicator *innovator* taking on 1 if the company introduced an innovation in the last three years. The *size class* variable is coded 1, 2, 3, and 4 if the company has less than 20 employees, 20 – 50, 50 – 250, and more than 250 employees, respectively. This ordinal indicator has the same content as the continuous variable *number of employees* and yet both variables are used jointly in practical applications. Usually it is argued that the categorical indicator could pick-up some non–linearities between *turnover* and *number of employees*[10].

The estimation results are summarized in table 5. The OLS–estimates indicate that both variables *number of employees* and *size class* are significant. The dummy variable *innovator* is not significant. However, the indirect method obtains that the *number of employees* is not significant on a 5% significance level. The latent variable underlying the categorical indicator picks-up the probably non-linear relationship between *employees* and *turnover*. That means that a transformation of the variable *number of employees* enters the linear regression but there is no indication about the type of transformation (Kukuk, 1994). The coefficient of determination is small, indicating that the specification used does not describe the data well.

In a next step, we consider the transformed variables log(*number of employees*) and log(*turnover*) in the otherwise unchanged linear regression. The estimation results of this specification are given

[10] The usual approach would be to include 3 dummy variables for the different size classes. As argued above, the biases that occur this way are at least qualitatively comparable to just using one indicator. The results are also not affected whether we use the integer coding scheme or the conditional expectations discussed in section 2.

Table 5: Regression estimates for Turnover (standardized)

	OLS		Indirect Estim.	
Regressor	Parameter	t–value	Parameter	t–value
C	-0.105	-2.215	0.045	2.146
Employees (stand.)	0.234	12.580	0.243	1.607
Size class	0.058	3.431	0.085	3.549
Innovator	0.030	0.690	0.010	0.480
Obs.	2748		2748	
R^2	0.063		0.062	

in table 6, clearly indicating an increased coefficient of determination. The ordinal variable does not indicate an effect anymore, but the dummy variable *innovator* is almost significant. This influence is reduced in the indirect procedure where the binary variable is interpreted to be ordinal with an underlying latent variable. In this transformed specification, a high correlation between the latent variable associated with *size class* and log(*number of employees*) is encountered being close to 1, whereas in the first specification this correlation is of magnitude 0.2. This can be interpreted as employment no longer having a non-linear relationship on the log(*turnover*) variable.

Our experiment is extended once more to demonstrate how estimation with a limited dependent variable could be performed. For a given vector of $\boldsymbol{\beta}$, the data generating process is simulated. Assume that our dependent variable is for instance binary, then this would be simulated as well. In the second step of the indirect method, an optimization is performed relating changes in the estimates to infinitesimal changes in $\boldsymbol{\beta}$. However, in this context a observational equivalence is likely to occur since for very small changes in $\boldsymbol{\beta}$ in finite samples, it is very likely that there is no change in the binary variable, implying that there is no change in the minimizing criterion. In the case of a continuous dependent variable, this problem does not occur since changes in the parameter vector result in changes in the simulated dependent variable and hence in the

Table 6: Regression estimates for Log(Turnover)

	OLS		Indirect Estim.	
Regressor	Parameter	t–value	Parameter	t–value
C	5.525	75.272	5.626	3.842
log(Employees)	0.988	26.399	0.980	2.683
Size class	-0.006	-0.0967	-0.034	-0.049
Innovator	0.124	1.908	0.105	1.502
Obs.	2748		2748	
R^2	0.623		0.970	

Table 7: Estimates for dichotomous turnover indicator

	Probit		Indirect Estim.	
Regressor	Parameter	t–value	Parameter	t–value
C	-2.440	-25.992	-0.0353	-8.093
Employees (stand.)	2.864	3.360	2.808	1.512
Size class	0.849	23.860	1.091	18.204
Innovator	0.173	2.487	0.0116	0.225

minimizing criterion.

One possibility to solve this problem is to use the latent interpretation of the probit model. The conditional expectation of the latent variable is modeled as a linear function of the regressor variables. The parameters of this linear function are estimated consistently using ML. Those parameters have up to a scalar factor the same interpretation as in the OLS case with the continuous dependent variable. This implies that replacing continuous regressors by their corresponding categorical indicators should lead qualitatively to the same biases. We exploit this fact in the indirect procedure, by comparing the probit estimates for the data at hand with OLS estimates as the auxiliary model for the simulated data. To be precise, not the whole assumed data generating process is simulated, since the dependent variable is not categorized. Instead, the continuously simulated values for the dependent variable are used in the auxiliary model which is a linear regression model. Therefore, usual optimization algorithms can be applied in the second step of the indirect procedure. To demonstrate this, the first specification using the levels of *turnover* and *number of employees* is used again, but this time *turnover* is transformed into a binary variable. The estimation results in table 5 now serve as the "true" model. The results of the probit estimation in table 7 indicate that the variable *innovator* has a significant effect which is not the case in the "true" model. The indirect method uncovers that *innovator* has no effect on latent *turnover* and also that *number of employees* has to be transformed to enter linearly in the conditional expectation of the latent dependent variable.

6 Discussion and Outlook

The experiments performed in this paper show that the indirect estimation procedure is a useful tool to test the influence of a latent variable in a linear regression approach although only categorical observations for that variable are available. Usually, it is the latent variable that is of main interest as shown in the last section. The *size class* indicator is used as a regressor but the results are usually interpreted as if observations for *number of employees* were used. The latent variable as a regressor variable is a natural extension of the concepts used in limited dependent variable models.

The additional assumptions necessary to apply the indirect estimation seem quite robust against violations as shown with the real data experiments. It is obvious that the data used violate the model assumptions. However, the results are satisfying in the sense that signs and significances of the “true model” are estimated correctly.

Another important result is that the method can be used even if the dependent variable is not measured directly. In this case, it is suggested not to simulate the whole data generating process, but instead to leave the simulated data of the dependent variable in its metric form. This implies that we apply two different auxiliary models, one for the data at hand and the other one for the simulated data. However, it is required that both auxiliary models estimate the same parameters consistently. This experiment shows that the indirect estimation procedure could be a promising way of also handling more complex models like probit models for panel data or simultaneous probit/logit models. In those models, the biases discussed in this paper are also likely to occur if only ordered observations are available for latent regressor variables. However, the latent models can be simulated as shown. The proper choice of an auxiliary model will be a crucial factor to estimating the parameters of interest efficiently.

References

Berenson, M.L. and D.M. Levine, (1992), *Basic Business Statistics: Concepts and Applications*, 5. Auflage, Annotated Instructor’s Edition, Prentice Hall, Englewood Cliffs, NJ.

Bierens, H.J. and J. Hartog, (1988), Non-linear Regression with Discrete Explanatory Variables, with an Application to the Earnings Function, *Journal of Econometrics, Vol. 38, No. 3*, 269-299.

Brillinger, D.R., (1982), A generalized linear model with Gaussian regressor variables, in: Bickel, J.,Doksum, K.A. and J.L. Hodges (Hrsg.), *A festschrift for Erich L. Lehmann*, Woodsworth International Group, Belmont,CA.

Browne, M.W. and G. Arminger, (1994), Specification and Estimation of Mean- and Covariance-Structure Models, in: Arminger, G. Clogg, C.C. and M.E. Sobel (Hrsg.), *Handbook of Statistical Modelling for the Social and Behavioral Sciences*, Plenum Press New York and London, 185-251.

Fang, K.-T., Kotz, S. and K.-W. Ng, (1990), *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London/New York.

Gallant, A.R. and G. Tauchen, (1996), Which Moments To Match?, *Econometric Theory, Vol. 12*, 657-681.

Gourieroux, C., Monfort, A., Renault, E. and A. Trognon, (1987), General Residuals, *Journal of Econometrics, Vol. 34*, 5-32.

Gourieroux, C., Monfort, A. and E. Renault, (1993), Indirect Inference, *Journal of Applied Econometrics, Vol. 8*, 85-118.

Hsiao, C. and D. Mountain, (1985), Estimating the Short-Run Income Elasticity of Demand for Electricity by Using Cross-Sectional Categorized Data, *Journal of the American Statistical Association, Vol. 80*, 259-265.

Jöreskog, K.G., (1990), New Developments in LISREL: Analysis of Ordinal Variables using Polychoric Correlations and Weighted Least Squares,*Quality and Quantity, Vol. 24*, 387-404.

Kao, C. and J.F. Schnell, (1987), Errors in Variables in the Multinominal Response Model,*Economics Letters Vol. 25*, 249-254.

Krämer, W., (1980), Finite Sample Efficiency of Ordinary Least Squares in the Linear Regression Model With Autocorrelated Errors, *Journal of the American Statistical Association, Vol. 75*, 1005-1009.

Kukuk, M., (1991), *Latente Strukturgleichungsmodelle und rangskalierte Daten*, Hartung–Gorre, Konstanz.

Kukuk, M., (1994), Distributional Aspects in Latent Variable Models, *Statistical Papers, Vol. 35*, 231-242.

Lee, S.-Y. and M.-L. Lam, (1988), Estimation of Polychoric Correlation with Elliptical Latent Variables, *Journal of Statistical Computation and Simulation, Vol. 30*, 173-188.

Li, M.M., (1977), A Logit Model of Homeownership, *Econometrica, Vol. 45*, 1081-1097.

Licht, G., C. Hipp, M. Kukuk and G. Münt, (1997), *Innovationen im Dienstleistungssektor. Empirischer Befund und wirtschaftspolitische Konsequenzen*, Nomos Verlag, Baden–Baden.

McIntosh, J., F. Schiantarelli and W. Low, (1989), A Qualitative Response Analysis of UK Firms' Employment and Output Decision, *Journal of Applied Econometrics, Vol. 4*, 251-264.

Nerlove, M., Ross, D. and D. Willson, (1993), The importance of seasonality in inventory models: Evidence from business survey data, *Journal of Econometrics, Vol. 55*, 105-129.

Olsson, U., (1979), Maximum Likelihood Estimation of the Polychoric Correlation Coefficient, *Psychometrika, Vol. 44*, 443-460.

Pearson, K., (1901), Mathematical Contributions to the Theory of Evolution. VII. On the Correlation of Characters not Qualitatively Measurable, *Philosophical Transactions of the Royal Society of London, Series A, Vol. 195*, 1-47.

Ronning, G., (1991), *Mikroökonometrie*, Springer-Verlag, Berlin.

Ronning, G. and M. Kukuk, (1996), Efficient Estimation of Orderes Probit Models, *Journal of the American Statistical Association, Vol. 91*, 1120-1129.

Ross, D.R., (1987), Estimating Linear Models with Categorical Indicators, Working Paper, Williams College, Williamstown, MA.

Ross, D.R. and K.F. Zimmermann, (1993), Evaluating Reported Determinants of Labour Demand, *Labour Economics Vol. 1*, 71-84.

Ruud, P.A., (1986), Consistent Estimation of Limited Dependent Variable Models Despite Misspecification Of Distribution, *Journal of Econometrics, Vol. 32*, 157-187.

Schepers, A. and G. Arminger, (1992), *MECOSA: A Program for the Analysis of General Mean- and Covariance Structures with Non-Metric Variables, User Guide*, SLI-AG, Frauenfeld, Switzerland.

Stoker, T.M., (1986), Consistent Estimation of Scaled Coefficients, *Econometrica, Vol. 54*, 1461-1481.

Theil, H., (1971), *Principles of Econometrics*, North-Holland, Amsterdam.

Yatchew, A. and Z. Griliches, (1984), Specification Error in Probit Models, *The Review of Economics and Statistics*, 134-139.