Turku Center for Welfare Research Working Papers
on Social and Economic Issues 08/2017

A General Method for Comparing Probit- and Logit-models with Single and Multilevel Data

In memory of Jukka Veilahti

Antti Veilahti

Turun yliopisto
University of Turku

Åbo Akademi
University

15.08.2017

A General Method for Comparing Probit- and Logit-models with Single and Multilevel Data

Antti Veilahti

Abstract

The paper proposes a method for overcoming the so-called latent scale-problem that prevents nested logistic and probit models from being compared. This allows us to decompose direct and indirect effects for binary outcomes. Our solution is based on an explicit construction of a latent propensity behind a given binary variable. The method is validated based on both simulated and the European Social Survey data. It is more accurate and easier to interpret than the previously available methods. Furthermore, it is the only method allowing us to compare mixed binary models: the so-called *y*-standardisation method, for instance, is not suitable for multilevel data because there is no global scale parameter applicable to both fixed and random effects. Finally, the paper concludes that the reason why nested binary regression models are not comparable is not related to 'unobserved heterogeneity', like Mood (2010) suggested, but it reflects the structure of the observed model.

Keywords: Logistic regression, mixed models, latent variable, decomposition

Introduction

One of the most topical methodological debates in quantitative social sciences relates to the comparison of categorical regression models. In the case of continuous variables, linear regression models can be compared in order to distinguish between direct and indirect effects. For instance, to what extent social class directly regulates income and to what extent it is mediated by the level of education. Likewise, it is possible to compare error variances across models so as to evaluate how much of the variation of income overall is explained by class or education (Prairie 1996). Such direct comparisons are not suitable for binary outcomes such as unemployment or the occurrence of a disease (e.g., Karlson, Holm, and Breen 2012).

This problem has been long known (Amemiya, 1981; Bollen, 1989: 238–246; Gail, Wieand and Piantadosi, 1984; Long, 1983: 49–52; Wooldridge, 2002: 470–472), but it used to be omitted by most sociologists. It then came as a striking news when Carina Mood (2010) reiterated the issue few years ago. Technically, the problem bears witness to the fact that categorical regression models appear to be specified only up to an 'unobserved' parameter (ibid.: 67) and which depends on the chosen set of predictors. However, we will argue that in acutality the size of the error is actually a structural property of the larger model: unlike what Mood argued, it is an 'observable' aspect of that model.

Such discrepancies are easy to understand intuitively: as there are more predictors in a larger model, the intercept group becomes more specific and the level of heterogeneity internal to that group is reduced. This results in more specific likelihood estimates that, for linear regression models, would result in the diminishing size of error variance. Unfortunately, in categorical models the error term is not available and the change of scale cannot be directly accessed.

There is another, more precise and sleek description of this issue that forms the basis of our approach. Indeed, it can be shown that binary regression models are equivalent with linear regression models, if we replace the binary outcome y with a suitable latent propensity y^* so that $y^* > 0$ whenever $y = 1$ (Long 1997: 47–50). Instead of modelling y, we can use linear methods to modelling y^*. While it is debatable whether it is plausible to assume the existence of such a propensity in the first place—what would it mean to say that someone is, say, more 'manly' than someone else—the idea is that the binary variable looses some information: it is unable of differentiating between those who are close to being men and those who are women by margin. If such a propensity y^* did exist, however, we could use this latent variable to observe the relative level of heterogeneity between the two models.

In this paper, we will generate this missed 'information' by hand. This approach might sound rather artificial, but we will show that the results do not depend on the

contents of this 'information'. Therefore, we will adopt a pragmatic position and claim that such a propensity y^* exists at the level of data—regardless of its meaning or existence at the level of social reality: there is no need to debate on whether there is a trait like 'manliness' but we are only dealing with constructs related to the particular set of data.

It is crucial that the binary and latent linear models agree as long as the model satisfies the suitable conditions regarding the error distribution. The only requirement is that y^* is scaled by a constant factor so that the error term is normally (or logistically) distributed with variance fixed at 1 (or $\pi^2/3$ for logistic models). The variance can be fixed because only the sign of y^* is actually observed (whether or not $y^* > 0$) and multiplying y^* by any positive number results in the same outcome y.

This equality between the binary and latent models under those conditions also stands for the very source of error. Indeed, when considering models involving a different number of predictors, *two different latent propensities* need to be used in order for them to satisfy the distributional conditions. Both of them need to be scaled according to their own error terms, and that is the precise reason why the corresponding binary models are not comparable (Mood 2010). There is also another problem identified by Karlson et al (2012): as long as the dropped predictors do not follow a suitable distribution (e.g. normal), not only the scale of y^* is altered but also its shape. This problem is avoided by our method, which uses a shared variable y^* across nested models. Even if the error distribution of the submodels does not itself satisfy the distributional conditions, this is the case always when comparing linear regression methods, which is a standard approach. Furthermore, the discrepancies will turn out to be much less severe than those involved in the comparison of binary models.

In this study, we will first explain our construction of an actual propensity y^*. We will then empirically validate the method in single- and two-level settings: we will show that the method yields deterministic results which are more credible than those produced by other methods, like the y-standardisation approach. It turns out that our method is particularly accurate for single level data whereas the random effects are slightly biased due to the shrinking of the maximum likelihood estimates.

1 Constructing a Latent Propensity

Instead of partaking the ontological debate on whether there actually is a latent propensity for a given binary variable y, we are dealing with a finite set of data. All constructs that we suggest exist as part of this data—not the reality to which it refers. It is an empirical question whether they result in reliable or useful estimates.

The following construction will be presented in the context of hierarchical probit- and logit-regression models, but the method is directly applicable to single level data by omitting the random components γ and u.

A mixed logistic regression model is given as

$$\ln \frac{p}{1-p} = x\beta + r\gamma + u, \tag{1}$$

where p is the expected value of y given x, r, γ and u. Similarly a probit-model is given by

$$\Theta^{-1}(p) = x\beta + r\gamma + u, \tag{2}$$

where Θ^{-1} is the inverse of the normal cumulative probability distribution. For simplicity we have not written out the indices of the different predictors or interactions: x and r are (horizontal) *vectors* of fixed and random variables, β and γ are scalar vectors, and u is the sum of all higher level residuals (there can be more than two levels).

To those readers new to mixed modeling, these models are similar to single level models except for the fact that for some effects (r) the coefficient is allowed to vary across contexts (γ), as does the constant term (u). In addition, we assume that the expected values of u and γ vanish and that these are constant within any single context. Depending on whether the random covariance matrix is structured, further conditions over the covariances between u and the different components of the random vector γ.

It can be shown that if there is a latent propensity y^* for which the multilevel regression model

$$y^* = x\beta^* + r\gamma^* + u^* + e^* \tag{3}$$

satisfies the standard conditions (e^* is either normally or logistically distributed with variance equal to 1 or $\pi^2/3$), then the respective binary model (1) or (2) agrees with the latent model (3). This means that $y^* - e^*$ is either $\Theta^{-1}(p)$ or $\ln \frac{p}{1-p}$. Given that $y^* > 0$ if and only if $y = 1$, the latter is equivalent with

$$e^* > -(x\beta + r\gamma + u). \tag{4}$$

Error Term

Instead of assuming the existence of such y^*, the idea is to start instead from a binary variable y and explicitly create a suitable variable e^* so that its conditional distribution is normal (or logistic) and that it satisfies the aforementioned equivalence. Given that we already know the binary model on y, we can then use such a

variable to construct

$$y^*_{construct} = x\beta + r\gamma + u + e^*,$$

which satisfies all the requirements for being a suitable latent propensity on y.

Such a variable e^* can be constructed in the following way: for each case, calculate the expected conditional probability p and dividing the normal (or logistic) distribution into two sides at a suitable point, which is to be determined by the respective binary model. Then pick a value from the appropriate side of that distribution.

In practice, we thus first need to calculate the conditional probability by conducting the probit- or logit-analysis. For each observation, the probability p that $y = 1$, and thus that the error should be in the upper part of the suitable distribution, is the inverse of the link function of the model-estimate $x\beta + r\gamma + u$. The error e^* is then reflected by a cumulative probability value q so that $q <= p$ if $y = 0$ and $q > p$ otherwise. In both cases, we can generate q from the uniform distribution and apply the link function $\Theta^{-1}(q)$ or $\ln \frac{q}{1-q}$ in order to acquire a suitable candidate for e^*.

It is easy to see that the variable e^* constructed in this way satisfies the required distributional conditions. Namely, for each observation the probability that $q < p$ is p and thus the conditional distribution of $q \mid x, r, i$ is the uniform distribution between 0 and 1. The resulting distribution is thus normal (or logistic) given the used link function, and its variance is independent of x, r and i, guaranteeing homoscedasticity. The construction also ensures that the inequality (4) is satisfied if and only if $y = 1$.

Adjustment

In theory, the conditional distribution of e^* for given x, r and context i satisfies the required conditions. However, the actual variable e^* is based on a finite random sample. Although it is reasonably close to the suitable distribution and is a latent propensity of y, it is not entirely independent of x, r and the context i. To adjust e^*, we can start by modelling e^* itself so that

$$e^* = x\beta_{adj} + r\gamma_{adj} + u_{adj} + e_{adj,raw}. \tag{5}$$

The error term $e_{adj,raw}$ satisfies the required distributional conditions while being independent of x, r and u. Particularly in the multilevel case, where the relative variance of the different components plays a more crucial role, it is mandatory to further adjust the error term by a suitable scalar constant in order for the variance of e_{adj} to equal to the theoretically expected value ($\pi^2/3$ in the logistic case or 1 in

the case of probit-models).

These adjustments are so small that $e_{adj} > -(x\beta + r\gamma + u)$ if and only if $y = 1$ still holds in about 99,8 % of the cases. While the simulated propensity

$$y^*_{adj} = x\beta + r\gamma + u + e_{adj}$$

thus fails to reproduce y unanimously, this is not really a concern because the actual value of y does not affect the model when β, γ and u are already known. Even so, by constructing e_{adj} similarly ten times and choosing the one with the lowest number of errors, it was possible to reduced the number of discrepancies by half.

Some Further Remarks On the Construction of y^*

It was an essential part of the introduction of multilevel models in the 1980's to realise that the maximum likelihood estimates of the random parameters are more conservative than the mean values, making the random parameters smaller than the observed group averages. For instance, in the case of the variance component model where there is only the higher level residual u and when normal errors are assumed, the actual estimate of the residual term u is

$$(y^* - x\beta)\frac{\mathrm{var}(u)}{\mathrm{var}(u) + \mathrm{var}(e)/n_i},$$

where n_i is the number of observations in the context i (cf. Rashbash et al. 2015: 39). Part of the observed higher level differences are then attributed to individuals, as reflections of random variation. This violates the assumption of the independence of levels, however. If we use a mixed model when constructing e_{adj}, it is not actually independent of higher level effects. Similarly, the retrospective model estimate of e_{adj} is not the same as the original construct e_{adj}.

To avoid the first problem, the equation (5) can instead be composed by considering both $r\gamma$ and u as fixed effects. This is done by using single level linear regression instead of multilevel regression, by incorporating the context as a categorical predictor (so that each context is associated with a different constant) and including the fixed interactions of the original random effects with context. By constructing e_{adj} this way, it appeared that both the fixed and random parameters of the model are independent of e_{adj} at least to six digit accuracy (i.e. relative errors are lower than 0,0001 %).

However, the shrinking of the random parameters also affects the initial model (2) and, again, when the produced latent variable y^* is modelled. This makes the random parameters produced by our approach less reliable than the fixed part of the model. We examined whether y^* itself instead of the error term could be constructed

by a fixed effects models, but the results were more ambiguous than when y^* is based on a mixed model, regardless of the adjustment. However, it will turn out that the underestimation is reasonably consistent, still providing a suitable basis for comparison across nested models, when the size of contexts is sufficient.

There are two other concerns that need to be addressed. The first issue relates to the set of predictors x and r used when constructing the latent propensity. Above, we suggested that y^*_{adj} has to reflect the full model, incorporating all predictors used while comparing models. As we will demonstrate below, this will give reasonable results. By contrast, we also tested the possibility that we could use some smaller set of predictors. Because y^* (almost) fully reproduces y, there is no loss of information whichever way y^* is constructed. However, this does not mean that a linear model on y^* should always reflect y adequately: the additional predictors on y would be exhibited only to a reduced extent (cf. Cox et al, 1992).

Second, we examined whether the results would be similar if we did not require the inequality (4) but e_{adj} would be independent of the observed outcome y. As long as we construct y^* based on all considered predictors, this does not appear to affect the fixed part of the model, which is further evidence to the fact that the latent scale factor is actually a property of the model (2) instead of being related to the 'unobserved heterogeneity' of the outcome. However, due to the shrinking of the random components, the fact that e_{adj} reflects y makes the random part more reliable.

Different Ways of Rescaling Binary Models

The basic motive for the above construction is that we can use the constructed variable y^*_{adj} instead of y as a basis of approximating the regression coefficients β, γ and u. However, it has been suggested that we could instead compare binary models themselves after rescaling the submodel by a suitable global parameter. Winship and Mare (1984) suggest the submodel to be scaled by a factor

$$\sqrt{\mathrm{var}(y^*)/\mathrm{var}(y^*_{sub})}, \tag{6}$$

where $y^*_{sub} = x'\beta' + r'\gamma' + u' + e'$ is a latent propensity on y corresponding to the submodel (e' is normal and independent with variation equal to one). Alternatively, we can use the full construct y^* as a basis of a submodel

$$y^* = x'\beta'_{lat} + r'\gamma'_{lat} + u'_{lat} + e'_{lat} \tag{7}$$

and use

$$\sqrt{\mathrm{var}(e'_{lat})/\mathrm{var}(e_{adj})} = \mathrm{st.dev}(e'_{lat}) \tag{8}$$

as the scaling parameter. It is easy to show that in the single level case these two parameters agree if and only if the rescaled binary models agree with those produced by the linear submodel (7), that is, when the omitted terms are normal. This is because they are absorbed in the error term e'_{lat}, which is otherwise not normal.

Even so, because we are actually scaling the fixed part of the model, it is more appropriate to use

$$\sqrt{\text{var}(x'\beta'_{lat})/\text{var}(x'\beta')} \tag{9}$$

instead. In the single level case, the random terms are omitted and this can be written as

$$\sqrt{\text{var}(y^* - e'_{lat})/\text{var}(y^*_{sub} - e)} = \sqrt{\frac{\text{var}(y^*) - \text{var}(e'_{lat})}{\text{var}(y^*_{sub}) - 1}}.$$

The x-standardised coefficient (9) agrees with (6) or (8) if and only if (6) and (8) are actually equal[1]. This only occurs when the omitted predictors are normal (or logistic), that is, when the rescaled binary model agrees with the linear submodel on y^*.

In the multilevel case, we can also use the the coefficient

$$\sqrt{\text{var}(x'\beta'_{lat})/\text{var}(x'\beta')}$$

for fixed effects, but it does not agree with the one reflecting the fitted parts $\sqrt{\text{var}(y^* - e'_{lat})/\text{var}(y^*_{sub} - e)}$. For random coefficients it would be best to use

$$\sqrt{\text{var}(r'_{lat}\gamma' + u'_{lat})/\text{var}(r'\gamma' + u')}$$

or even a specific scaling parameter for each random component. Unless it agrees with (9), it is not meaningful to rescale mixed binary models as a basis of comparing random effects.

There are now three questions that the rest of this paper seeks to answer: to what extent do these different scaling-parameters agree? Second, which one of them is the best method for rescaling nested binary models? And third, when there is disagreement between them, is it more appropriate to compare linear submodels on y^* rather than optimally scaled binary submodels?

2 Research Design

In the rest of this paper, we will evaluate the reliability of the method empirically, seeking to answer the three questions asked above. In particular, we will compare the results of the latent linear model with differently scaled binary submodels.

Data

We analysed the method by using both random data and the sixth round of the European Social Survey (2012). In the latter set of data, NUTS2-units as the higher level units, each consisting of from 5 to 2380 respondents and with 36 402 altogether. The European Social Survey (2012) data was used to assess the method in practical situations, whereas random data was used to assess the applicability of the method with variables reflecting various distributions and particularly with variables independent of other terms so that the results could be compared with theoretical expectations.

With simulated data, we used 1000 random observations for single level models. It was the intent to try the method with such a limited set of data to see whether it is still determinist. In the multilevel case, we used 3000 or 30 000 random observations from 50 or 500 contexts (with the average context size of 60 or 600). In the latter case, we analysed an outcome y which was based on an actual latent variable $y = x_1 + \ldots + x_8 + x_1 \cdot (x_9 + x_{10})$, where $x_1, \ldots, x_5$ were lower level variables and $x_6, \ldots, x_10$ depend only on the context. For single level models, all variables were defined at the lowest level. They were chosen to reflect binary, normal and uniform distributions.

Procedure

The results are expressed for probit-models, but logistic models were found to produce a comparable level of error. For both sets of data, we chose various initial models and then compared different submodels of y^* with different binary submodels on y. We varied the five predictors used for the full model and the number and set of predictors included in the submodel. For multilevel models, we also altered whether the random effect was incorporated as part of the submodel.

In each case, the different scaling factors proposed in the previous chapter were calculated, while the corresponding fixed and random terms based on the rescaled binary models were compared with those produced by the linear submodel of y^*. Both approaches to error adjustment were tried. The fixed coefficients were compared both on average and by calculating the sum of squares

$$\sum_i (\beta_{i,lat} - scl \cdot \beta'_i)^2$$

for different scaling factors scl.

In the context of the ESS data, in contrast, we sought to examine the usefulness of the model in more complex situations. The comparisons were conducted in three different settings depending on the structure of the full model. Each combination

of outcome and predictors was used as a basis of comparing 11-16 binary and linear models and with 64 combinations of variables. In all the three settings, the outcomes and variables were chosen among those expressed in Table 1 (p. 10).

The first two settings were based on *variance component models* with no random effects. Each latent variable was constructed based on three lower level effects, one of which was binary and the others continuous, and one continuous higher level effect together with all two- and three-way interactions. In these two settings, the binary and simulated models were then compared for 16 different subsets of these 15 fixed effects or interactions. Both settings then resulted in 1024 pairs of submodels. It was our intent to test whether we could include all possible combinations of variables, examining whether the method is robust enough to handle a reasonable level of complexity.

In the third setting, the aim was to examine the behaviour of the *random effects* in practical situations. The models included a random parameter that, depending on the model was either continuous or binary. We particularly asked whether replacing a random effect by a fixed one would still produce reliable results. In this setting, 11 submodels were used resulting in 704 pairs of submodels altogether.

Table 1: Variables used in the analyses

role	variable
outcomes	anxious
	voted
	boycott
	trust eu
fixed effects	father's socioeconomic status
	lr-scale
random effects	neuroticism
	hinctnta
	gndr
	chldhm
higher level effects	GDP per capita
	popularity of tertiary education

3 Results I: Single Level Models

Above we constructed a latent variable y^*_{adj} whose sign depends on a given binary outcome y—in 99,8 % of cases at least—and which satisfies the distributional conditions for a suitable set of predictors. Theoretically, a mixed model on y^*_{adj} should then reproduce the binary regression model on y.

In the case of single level models, this appears to be the case: the full models are produced with at least five digit accuracy, meaning that all of the β-coefficient estimated based on y^*_{adj} differ from the binary coefficients on y by less than 0,0001 %. As a result, the binary submodel also agrees with the corresponding latent model constructed on the basis of this binary submodel

$$y^*_{sub} = x'\beta' + e', \tag{10}$$

which differs from the one based on the original latent construct

$$y^* = x'\beta_{red} + e_{red}. \tag{11}$$

The latent error term e_{red} incorporates part of the omitted terms that are included in x but not x' and r'. If these variables are normal, then (11) should agree with (10).

Previously we discussed the best way of reparametrising the latter equation so as to compare the two. If (6) and (8) agree, there is no difference between the three scaling parameters: $1/\sqrt{\text{var}(e_{red})}$, $\sqrt{\text{var}(y^*)/\text{var}(y^*_{sub})}$, and $\sqrt{\text{var}(x'\beta_{red})/\text{var}(x'\beta')}$. In addition, we considered two additional scaling parameters constructed by hand: a weighted average over the quotients β_{red}/β' and the quotient of the sums of the norms of the two arrays of coefficients.

The Most Appropriate Scale Parameter

We examined the accuracy of these parameters when normalcy of the omitted terms is not assumed. This was done by considering the sums of squares comprising the difference between the rescaled coefficients and the original ones

$$\sum(\beta'_{rescaled} - \beta_{red})^2.$$

In 384 sets of simulated data we analysed, the measure of the fitness of the scale parameter was the lowest for $\sqrt{\text{var}(x'\beta_{red})/\text{var}(x'\beta')}$, averaging at .0007, and with the maximum value of .0117, as illustrated in Table 2. This suggests that even in the worst case the difference of the coefficients depending on the method should be much less than 10 %, as long as the proper scale-parameter is used. As a brief note, the errors indicated in Table 2 increase linearly as a function of the variance of $x\beta$, indicating that the errors are limited relative to the size of $x\beta$.

Based on Table 2, comparing the different approaches to reparametrising the model, it is obvious that $\sqrt{\text{var}(x'\beta_{red})/\text{var}(x'\beta')}$ is preferrable. Only the parameter based on y-standardisation has been previously accessible, however, and this has further contribute to the biases of the binary approach.

Table 2: Sum of squares of the differences of the β-coefficients in the rescaled binary model and the respective submodel on y^*, simulated single level data.

scaling parameter	mean difference	sd	min	max
$\sqrt{\text{var}(x'\beta_{red})/\text{var}(x'\beta')}$	.0007	.0014	0	.0117
$1/\sqrt{\text{var}(e_{red})}$	.0011	.0018	0	.0129
$\sqrt{\text{var}(y^*)/\text{var}(y^*_{sub})}$	.0017	.0029	0	.0170
sum of the norms	.0012	.0063	0	.0626
weighted average	.0032	.0272	0	.2712

Which Approach is the Best?

Yet the models based on y^* and y^*_{sub} differ not just by a scale-parameter but also structurally, as suggested by the differences identified above. We examined whether the most appropriate rescaled binary model or the submodel based on the full latent parameter y^* would be more credible basis for decomposing the original model. This was done by adjusting the predictors included in the submodel so that they would be independent of the other parameters. In this case, it is a mathematical fact that the respective coefficients in the submodel should reflect their extraction in the full model.

This appeared to be the case when comparing the full model with the latent submodel on y^*: the average difference of the β-coefficients was found to be negligible (lower than 0.001 %). In contrast, the rescaled binary model appeared to underestimate the coefficients by almost 2 % on average (Table 3), suggesting that the model violates the expected structural composition. The latent submodel is thus more appropriate than the binary one when decomposing the direct and indirect effects, at least when independent effects are involved.

Table 3: Extraction of independent effects in binary submodels relative to the full model, %.

distribution	mean difference	sd	min	max
normal	100.2	1.7	96.1	103.9
uniform	100.0	1.4	96.1	103.8
binary	100.1	1.3	97.2	103.9

4 Results II: Mixed Models

In the single level case, the method is deterministic and provides much more accurate estimates for independent effects than the rescaled binary models. In the multilevel case, the results are similarly deterministic, with at least six digit ac-

curacy $(0,0001\%)$ when the error term is adjusted based on a fixed model. By contrast, adjusting the error based on a random effect model resulted in slight variation in the estimate as summarised in Table 6. However, in cross-model comparisons such differencies are diminished because the same error term is used for both of the two nested models, making the comparisons reasonably accurate even with smaller contexts.

Table 4: Standard deviation of the coefficients when the error adjustment is based on a mixed model with the average size of context 60 and 600, %.

effect type	small contexts sd	large contexts sd
lower level β-coefficient	< 0.2	$< .003$
higher level β-coefficient	< 2.0	$< .04$
The random γ-coefficient	4.1	0.1
Change of γ across models	0.5	$< .01$
Higher level residual u	7.9	1.4
Change of u across models	1.2	0.3

Full Model

As with single level models, we started by comparing the full binary model with the full model on y^*. The results were slightly less unanimous (see Table 5). With fixed error adjustment which guarantees deterministic results, the latent model reproduces the lowest level effects reasonably well, with the standard deviation of the difference being less than 0.2 % for all tested distributions. In the case of mixed error adjustment, in contrast, the difference was about 1 %, fifth of which is explained by the indeterminacy of the approach and the rest by structural differencies. As with single level models, with the overall differences of the β-coefficients is proportional to the standard deviation of the $x\beta$-term.

Table 5: Difference of the full binary and linear models for different effect types, %.

	small contexts, fixed		large contexts, fixed		small contexts, mixed		large contexts, mixed	
effect type	mean	sd	mean	sd	mean	sd	mean	sd
Lowest level fixed effect	100.0	0.22	100.0	0.01	100.0	0.1	100.0	< 0.01
Fixed effect behind a random effect	99.9	0.58	100.1	0.38	100.0	0.03	100.0	0.02
Higher level fixed effect	100.3	2.8	100.6	1.8	100.0	1.8	99.8	0.6
Constant	99.7	1.5	100.0	0.04	100.6	1.4	100.0	0.02
Residual u	78.2	24.8	113.8	19.8	91.5	16.1	101.7	10.2
Random effect γ	91.4	14.2	110.1	10.4	98.3	9.3	100.6	6.7

The differences indicated in Table 5 result from the fact that the maximum likelihood estimates of the random terms are subject to shrinking: when constructing the binary model, this shrinking occurs only once. By contrast, the latent construct y^* itself reflects the originally shrank model, while the linear model on y^* is then subject to shrinking the second time. Even so, the results were found to be more reliable than in alternative approaches that would avoid the effects of double shrinking[2].

Comparing Nested Models

The underestimation of the random parameters due to double shrinking is not necessarily a flaw, however, if the shrinking is consistent enough so that nested models can be meaningfully compared. While no previously valid method exists by which we could verify the validity of such comparisons entirely, we can look for indirect evidence. In particular, we can demonstrate that the results are much more reasonable than when comparing binary models directly.

In the binary case, the submodel sometimes gave much lower random parameters than the full model, which should happen only rarely. A similar phenomenon did not occur to a notable extent when comparing latent models. Moreover, the variance of the difference of the random components was much higher when comparing rescaled binary models instead of the linear ones. This suggests that rescaling binary models is not a meaningful approach in regard to the random effects: there is no global scale parameter that would allow the y-standardisation method to be extended to multilevel settings, but each random component instead requires its own scale parameter.

On the other hand, the higher level residual is closely connected to the higher level fixed effects used. When looking at a higher level effects that is independent of other effects, the linear submodel on y^* gives reasonable results with a standard deviation of 3 %, whereas the nested binary models would misrepresent the effect by 15 % on average (Table 6). Also in more complex settings, where there higher level effects are not independent, there occurs fewer outliers in the linear approach, making it a more credible candidate than the binary one.

The contrast between the binary and linear approaches is similarly apparent when looking at lower level fixed effects. The linear submodels were much closer to theoretical expectations, with 0.2 % accuracy, whereas the nested binary models discrepancies with the standard deviation as high as 1.6 %. For both methods, the accuracy was notably higher when increasing the size of contexts to 600, but the relative difference between the two approaches remained similar.

Table 6: Accuracy of the independent effects relative to theoretical expectations, %.

	fixed adjustment		mixed adjustment		binary model	
effect type	mean	sd	mean	sd	mean	sd
Lowest level fixed effect	100.0	1.0	100.1	1.0	100.0	3.5
Fixed effect behind a random effect	100.0	1.2	100.0	1.2	100.0	3.7
Higher level fixed effect	100.4	3.4	100.3	4.2	102.5	23.9
Higher level fixed effect (large contexts)	100.1	0.6	100.1	0.6	102.2	8.5
Residual term u	98.4	1.0	98.4	1.2	95.6	4.9
Random parameter γ	99.1	1.8	98.8	2.0	95.6	7.2

Against this background, we wanted to measure the difference between the two methods in more complex settings, where variables need not be independent. As

illustrated in Table 7, the differences are notable. In some respects, but not all, the rescaled binary model approaches the linear one when the size of contexts becomes substantial: this does not apply to the higher level effects or the random coefficients.

Therefore, unlike in the single level case, in the multilevel setting the rescaled binary models should not be used at all: the discrepancies are limited for the lowest level fixed effects similarly as with single level data, but not with higher level effects or the random parameters. The comparison also shows that in regard to the random parameters, the approach based on the fixed error adjustment is as biased as is the binary approach (Table 7).

Table 7: Difference between cross-level comparisons in the latent and binary settings, %.

	small contexts				large contexts			
	fixed		mixed		fixed		mixed	
effect type	mean	sd	mean	sd	mean	sd	mean	sd
Lowest level fixed effect	98.7	6.0	99.2	6.1	99.7	2.6	99.7	2.6
Fixed part of a random effect	98.0	9.6	98.4	9.6	100.9	10.9	100.9	10.6
Higher level fixed effect	105.2	7.1	102.9	7.2	99.5	24.6	102.1	24.6
Constant	96.1	2.4	96.0	2.4	97.3	6.2	97.3	6.2
Residual u	121.1	53.5	105.8	32.9	109.3	31.6	102.6	14.4
Random effect γ	101.1	11.7	99.7	11.9	100.1	3.1	99.9	2.5

Choosing the Right Approach to Error Adjustment

Both approaches to error adjustment have their own benefits. For smaller contexts, the choice depends on if we want to emphasise the determinacy of the approach and β-coefficients, or if we are instead interested in the higher level effects. When the size of contexts is large, the choice is less urgent. In general, we should only use the fixed error adjustment when seeking to decompose fixed effects. By contrast, when seeking to decompose the higher level effects or cross-level interactions, it seems more appropriate to utilise the mixed adjustment instead.

A Brief Comparison of the Linear and Binary Models with ESS Data

While the simulated data demonstrates that the linear approach gives much more credible results than the binary one, we examined the differences also in the context of the European Social Survey. In practical social scientific research, the higher level effects usually account to only a fraction of the overall variation, and we wanted to know whether the rescaled binary approach could have practical relevance in such applications.

The differences between the latent and rescaled binary approaches are again prominent, however (Table 9). It appears that continuous effects tend to be under-

Table 8: Relative difference of the different components of the rescaled binary submodels of y and the corresponding linear submodels of y^*, %

model type	mean	st. dev.	st. err.
constant	100.0	3.2	0.1
fixed effect	99.0	8.7	3.2
fixed part of a continuous random effect	101.6	5.3	0.2
fixed part of a binary random effect	97.8	16.8	0.6
interaction with a continuous random effect	102.6	11.3	0.8
interaction with a binary random effect	96.2	16.3	1.5
higher level fixed effect	98.4	9.4	0.5
cross-level interaction	100.2	9.9	1.7
higher level residual u	90.2	9.8	0.2
random effect γ	91.6	17.1	1.1

estimated by binary models, whereas binary effects are usually slightly exaggerated but also more variable. The average difference is also higher when a binary variable is involved in the submodel. Moreover, the binary model becomes less reliable when the number of omitted terms increases (Table 9).

Higher level effects and interactions are even less accurate. On average, the changes in the random parameters are underestimated by the binary model, but these differences are even more variable than those associated with fixed effects.

Table 9: Average difference of the β-coefficients between the linear and binary models, by type of the submodel (x is a lower level fixed effect, r_c a continuous random effect, r_b a binary random effect, and z a higher level effect), %

	difference in %				$\sum(\beta'_{rescaled} - \beta_{red})^2$	
type of the model	mean	sd	min	max	mean	sd
constant	98.5	5.4	81.3	120.6	.0019	.0028
x	99.2	1.8	94.5	103.0	.0005	.0008
r_c, x	99.5	1.8	94.0	103.2	.0098	.0360
$r_c, x, r_c \times x$	101.0	1.9	92.9	103.6	.0006	.0016
r_b, x	101.0	3.5	94.9	110.5	.0011	.0020
$r_b, x, r_b \times x$	101.0	4.4	92.5	13.6	.0036	.0086
z	99.6	5.1	86.3	129.8	.0026	.0050
z, r_c	101.1	3.9	98.0	117.6	.0007	.0016
z, r_c, $z \times r_c$	100.2	1.4	97.6	103.3	.0002	.0003
z, r_b	102.4	4.1	96.0	113.0	.0014	.0027
z, r_b, $z \times r_b$	101.8	4.4	95.7	112.5	.0014	.0026
r_c, r_b	99.1	7.6	71.7	116.6	.0027	.0099
$r_c, r_b, r_c \times r_b$	100.2	1.7	95.0	105.2	.0041	.0245
z, r_c, r_b	101.1	3.8	95.5	116.6	.0034	.0084
z, r_c, r_b, $z \times r_c$, $z \times r_b$	99.9	1.5	96.3	103.6	.0003	.0005
full model	99.8	1.2	97.2	103.0	.0002	.0005

5 Conclusions

Probit- and logit-models continue to dominate quantitative social scientific research. In this paper, we have proposed a solution to a problem that makes the direct comparison of such models erroneous (cf. Mood, 2010). In this paper, we have proposed a solution to this problem. In particular, we have demonstrated that it is possible to construct a continuous variable y^* which reflects a given binary outcome y and which we can then analyse by linear methods. This allow us to decompose fixed and random effects.

Our method is the only currently available method for decomposing fixed and random effects on binary outcomes for multilevel data. With single level data, the method also gives results that are much closer to theoretical expectations than the binary approaches like, say, the y-standardisation method (Winship and Mare 1984; Long, 1997), which compromises both the validity and reliability of cross-model comparisons.

Even in the single level context, the previous solutions are not just less accurate, but they have other limitations as well. For instances, Mood's (2010) attempt to use marginal effects is difficult to interpret, particularly if there are large differences between classes. In contrast, the KHB-method makes assumptions about the added predictor. Also, the KHB-algorithm in STATA often failed to function when comparing models involving a large number of variables, while our method handles such settings with ease.

In the y-standardisation approach, by contrast, the two compared models refer to different latent variables. Not only do they differ by a global scale parameter but due to different error distributions. Only when the omitted variables are normal (or logistic), this difference can be avoided.

We have proposed two alternative ways of operationalising the method. One of them gives deterministic results, whereas the other one yields more reliable higher level and random estimates. With a sufficient size of contexts, there is no notable difference between these two operationalisations, however, whereas the binary approach continues to yield biased results.

This suggests that there are structural reasons why the y-standardisation method, or any alternative rescaling method is not suitable for comparing mixed binary models. In particular, there is no global scale parameter but the fixed and random parts of the model should be scaled differently.

By contrast, comparing linear models on a given continuous variable allows us to use standard decomposition tools, including the relative change of R^2-values (cf. Prairie 1996). The constructed variable y^* satisfies all the conditions required for it to be used as a basis of such methods. Our only concern is that the linear

(sub)models on y^* should adequately reflect the original variable y. This assumption has been verified in the context of single level data as well as the fixed coefficients (β) for multilevel data. For random coefficients we have demonstrated that our approach is more reasonable than the direct comparison of binary models.

As one shortcoming, our method does not provide adequate statistical metrics, and leave it for future research to establish whether such metrics can be recovered from y^*. Even so, we can still consult the corresponding binary models in order to assess the statistical scales for fixed and random effects. In other words, we can transfer the linear coefficients back to the binary context by using the inverse scale parameter $1/\text{st.dev}(e'_{lat})$ and used the metrics provided by the binary model, or we can verify whether the effects are significant directly from the binary model. For single level models, these estimates are reasonably accurate, whereas for multilevel models their accuracy can be estimated based on Tables 6 and 8. Of course, this does not allow us to estimate the significance of the cross-model comparisons, but even for linear models there is no adequate basis for examining the change of individual effects statistically. Instead, only the change in the overall fitness of the model (information criterion) can be tested.

In conclusion, we have demonstrated three things: the previously used scale-parameter is not the optimal one. Second, in complex settings, our method gives more accurate results in comparing nested models than those that have been previously available. Finally, our results question the previous argument that the comparison of binary regression models should fail due to 'unobserved' heterogeneity within data (Mood, 2010). Instead, we have demonstrated that the choice of the error parameter e, which stands for this 'unobserved' heterogeneity, does not affect the results (with at least six digit accuracy): nested models are not comparable because of the structure of the observed models, and not because of what remains unobserved about them.

Acknowledgments

This research was supported by the Strategic Research Council of the Academy of Finland (decision number: 293103).

References

Amemiya, T. 1985. *Advanced Econometrics.* Cambridge, MA: Harvard University Press.

Bollen, K. A. 1989. *Structural equations with latent variables.* New York: John Wiley.

Breen, R., Karlson, K. B., and Holm, A. 2013. "Total, Direct, and Indirect Effects in Logit and Probit Models." *Sociological Methods & Research* 42(2): 164–191. 10.1177/0049124113494572

Gail, M. H., Wieand, S. and Piantadosi, S. 1984. "Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates." *Biometrika* 71: 431–444.

Karlson, K. B. 2015. "Another look at the method of y-standardization in logit and probit models'. *The Journal of Mathematical Sociology* 39(1): 29–38.

Karlson, B. K., Holm A., and Breen R. 2012. "Comparing Regression Coefficients Between Same-sample Nested Models Using Logit and Probit: A New Method." *Sociological Methodology* 42: 286–313.

Long, J. S. 1983. *Confirmatory factor analysis.* Newbury Park: Sage.

Long, J. S. 1997. *Regression models for categorical and limited dependent variables.* Thousand Oaks, London and New Delhi: SAGE Publications.

Prairie, Y. T. 1996. "Evaluating the predictive power of regression models." *Canadian Journal of Fisheries and Aquatic Sciences* 53(3): 490–492.

Rasbash, J., Steele, F., Browne, W. J., Goldstein, H., and Charlton, C. 2015. *A user's guide to MLwiN.* Centre for Multilevel Modelling, University of Bristol, UK. Retrieved March 28, 2017 (http://www.bristol.ac.uk/cmm/media/software/mlwin/downloads/manuals/2-36/manual-print.pdf)

Snijders, T.A.B. and Bosker, R.J. 1999. *Multilevel Analysis.* Newbury Park, California: Sage.

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

Wooldridge, J. M. 2002. *Econometric analysis of cross section and panel data.* Cambridge: MIT Press.

Notes

[1]This is based on the fact that if $a/b = c/d$, then $a/b = (a-c)/(b-d) = c/d$.

[2]In order to avoid double shrinking, we tried to use either when constructing y^*, or alternatively, when analysing y^* constructed as above. The resulting random parameters were then overestimated, however, and the discrepancies were generally higher. We also tried weighting the cases in the retrospective analysis of y^*, say, by counting each case 100 times so that the size of contexts would increase. This is recommended only with fixed error adjustment, however, because otherwise the random components appeared to be greatly overestimated. While increasing the reliability of the full model, this weighting did not enhance the reliability of the relative change in the random parameters in cross-model comparisons, moving the models further apart from the corresponding binary models.