Forecasting with Unobserved Heterogeneity

Matteo Richiardi *[a,b,c,d]

[a]INET, University of Oxford

[b]Nuffield College, Oxford

[c]University of Turin, Department of Economics and Statistics.

[d]Collegio Carlo Alberto & LABORatorio R. Revelli

May 24, 2014

Abstract

Forecasting based on random intercepts models requires imputation of the individual permanent effects, when direct estimates are not available. If current outcomes are observed, this involves sampling from conditional distributions, which might be computationally burdensome. I review alternative approaches, from assuming null individual effects in forecasting, to regression-based imputation of the individual intercepts, and show their shortcomings. I then present an algorithm for drawing individual permanent effects from a conditional distribution which only requires to invert the corresponding estimated unconditional distribution. The algorithm, labeled *Rank method*, solves the optimal assignment problem in linear problems and offers a good approximation in binary response models. It only requires matching two ranks and works in $N \log N$ time. It is useful in linear models with fixed effects, and in binary response models with fixed or random effects.

*Email: matteo.richiardi@unito.it.

Financial support from Collegio Carlo Alberto and Regione Piemonte within the research projects "Causes, Processes and Consequences of Flexsecurity Reform in the EU: Lesson from Bismarckian Countries" (Collegio Carlo Alberto) and "From Work to Health and Back: The Right to a Healthy Working Life in a Changing Society" is gratefully acknowledged. This work benefited from very helpful comments from Christopher Flinn, Gaetano Carmeci, Cinzia Carota. All errors and omissions however are my sole responsibility.

Keywords: forecasting, microsimulation, random intercepts, unobserved heterogeneity, computational complexity.

JEL codes: C15 (Statistical Simulation Methods: General), C53 (Forecasting Models; Simulation Methods), C63 (Computational Techniques; Simulation Modeling).

1 Introduction

In this paper I present a new method for assigning individual specific effects to a population when no estimates are available, at the individual level. This generally happens when the population to be simulated (the *population sample*) is different from the population on which the model has been estimated (the *estimation sample*). The method is relevant for projecting forward in time models with unobserved heterogeneity (UH). UH generally comes in the form of fixed effects or random effects ('random intercept') models, where the deviations from the conditional expectation function are specified as

$$e_{i,t} = \alpha_i + u_{i,t} \tag{1}$$

with α_i being the individual-specific effect, that is, permanent UH, and $u_{i,t}$ a random component. The method described here is useful in linear models with fixed effects, and in binary response models with fixed or random effects.

Linear and binary choice models differ in the amount of information that can be obtained about the (unobserved) random intercepts. In linear models, non-parametric estimates of the individual random intercepts can always be obtained.[1] If we are concerned with projecting forward in time the estimation sample, the estimated individual intercepts will simply be treated as additional covariates.[2] In binary choice models the individual effects are in general not separately identifiable, and one has to content with an estimate of their standard deviation in a random effects setting; only inconsistent estimates of the individual intercepts can be recovered with fixed effects estimation.[3] How to make out-of-sample predictions in such models is therefore a non-trivial problem.[4] More-

[1] In addition, a parametric estimate of the shape of their distribution is also obtained, under the random effects assumption.

[2] This is generally done automatically by standard statistical packages, e.g. with the `predict, xbu` post-estimation command in Stata, after a `xtreg` model is run.

[3] The inconsistency of the fixed effects maximum likelihood estimates of nonlinear models in finite samples originates from the incidental parameters problem (the fact that the number of parameters grows linearly with population size) and affects also the estimates of the other coefficients (Neyman and Scott, 1948; Moon et al., 2014). The conditional logit approach circumvents this problem by maximizing the likelihood of the observed outcome $\boldsymbol{y}_i = (y_{i,1}, y_{i,2}, \cdots, y_{i,T})$ conditional on $\sum_t y_{i,t}$, which does not involve the α_i.

[4] In Stata the `predict, pc1` post-estimation command after a `xtlogit, fe` fixed effects regression predicts the probability of a positive outcome conditional on one positive outcome within the group; no

over, in both linear and binary models an issue arises if the estimates have to be applied to a different sample, for which the individual effects are (by definition) unobservable.[5]

Such a situation is indeed common, for instance in dynamic microsimulation. Microsimulation models generally include different processes (like schooling, household formation, labor market transitions, retirement, etc.): it is quite unlikely that a single dataset exists with all the relevant variables so that it can be used both for estimation of all processes and as a basis for simulation. A more common situation is to estimate different processes on different datasets, and than apply the estimated coefficients to some initial population to be simulated forward in time. True, to simulate we need information on all the variables included in the empirical specifications, but this falls short from requiring the union of all the datasets used for estimation, as (i) we do not need the longitudinal dimension required for dealing with UH, (ii) we do not even need retrospective information if the empirical specifications only include first order lags, as the base year values will become lagged values in the first year of the simulation, and (iii) we might impute missing information from other donor datasets[6].

Even when the initial population coincides with the estimation sample, it is often the case that it needs to be expanded over time, for instance to include partners, offsprings, or immigrants. These new individuals entering the simulation might come with a previous history of outcomes as well: not only in the case of spouses and foreigners, but also of newborns. If this sounds bizarre, consider that many datasets register information only for individuals above a minimum age: for instance, EU-SILC have data only for those aged 16+. Given the wide coverage of the EU-SILC survey, this is a likely feeder of a microsimulation model for European countries. In this case, newborns enter the microsimulation at age 16, after having already experienced meaningful education and labor market choices/lotteries.

Finally, even with cohort models where the evolution of a single cohort of individuals is simulated through time, it is quite likely that the cohort is not observed from birth, so

similar commands are available after a random effects regression.

[5]In this case, the out-of-sample nature of prediction regards both time and the units of analysis.

[6]By converse, estimating the models on imputed data would impinge on the properties of the estimates (Rubin, 1976; Little and Rubin, 1987).

that individuals enter the microsimulation with a previous history of outcomes.

Four broad classes of solutions can be conceived to the problem of assigning a random intercept to each individual in the simulation sample. First, we can simply forget the problem, set all random intercepts to zero, and take into account only the observables in simulating the outcomes of interest. This is often done and, as we will see, can be justified on some grounds, in particular when the model is linear. In nonlinear models, however, setting the random intercepts to zero in the simulation introduces a non-negligible bias in the projections. Second, we can impute the missing variables in the simulation sample from their estimated counterparts in the estimation sample by means of a regression model where the estimated random intercepts are modeled as a function of the observable explanatory variables and the outcome variable(s). This however offers only a partial solution, as we shall see, since it distorts the distribution of the random intercepts; moreover, this approach is feasible only when the individual effects are separately identified in the estimation sample, that is when the model is estimated with fixed effects. Third, we can draw the individual effects from the estimated distributions of the random intercepts, conditional on the observed past outcomes. This is trivial in linear models with Gaussian random effects, less straightforward in linear models with fixed effects, where the fixed effects follow an arbitrary distribution, and in binary response models with random or fixed effects. Fourth, we can sample from the unconditional (paramdetric or empirical) estimated distributions of the random intercepts, and then assign to each individual the value for UH that best matches his observed past outcomes, among those sampled. Methods for solving this optimal assignment problem can be borrowed from the linear programming literature. However, they work in (third or fourth order) polynomial time: this might be an impediment in forecasting exercises that involve hundreds of thousands or even millions of individuals, as is common in dynamic microsimulation models (Li and O'Donoghue, 2013). Moreover, as we shall see in nonlinear models the optimal assignment solutions introduce an artificial correlation between the imputed individual effects and the covariates.

A critical review of these approaches is the first contribution of the paper: to the

best of my understanding, this is novel in the literature, with the exception of a brief overview in Panis (2003). The second contribution is the development of a new assignment algorithm —labeled the *Rank method*— which works in quasi-linear ($NlogN$, where N is the sample size) rather than polynomial time, and provides an optimal imputation of the individual effects in linear regression models. In such models, the Rank method is valuable when the estimates are performed with fixed effects, as it only requires sampling from the unconditional estimated distribution of the individual effects, which is simple no matter the shape of the distribution. Moreover, when applied to binary response models the Rank method provides a reasonable approximation of the optimal assignment solution, while reducing the artificial correlation between the imputed individual effects and the explanatory variables.

The paper proceeds as follows. Section 2 introduces two illustrative models that will be used throughout the paper —a continuous response linear model and a binary response latent variable model. The following sections discuss alternative approaches to the problem of assigning the unobserved individual intercepts: section 3 discusses whether and when simulating UH leads to better forecasts; section 4 shows the shortcomings of imputing the random intercepts *via* a regression model; section 5 describes the Bayesian solution to the imputation of the random intercepts, which involves deriving the conditional distributions and then sampling via the Inverse Transform (IT) method; section 6 explains how the problem is approached in a linear programming setting, leading to solutions which have computational complexity of polynomial order and forcing the correlation between the imputed random intercepts and the explanatory variables in the binary case. The final sections are devoted to the Rank method: section 7 presents the algorithm and shows that it solves the optimal assignment problem in linear models, while section 8 shows that the application of the Rank method to binary response models leads to an increase in the forecasting error with respect to the optimal assignment solution, but reduces the induced correlation problem. Section 9 concludes.

2 Empirical setting: linear and binary response models

As a basis for discussion, I consider two benchmark cases. Eq. (2) refers to a simple linear model, where the (observed) continuous response is y^*. If we assume, by converse, that y^* is a latent (unobserved) variable, and that only a discrete 0-1 outcome y can be observed, we get the standard binary response model of eq. (3).

$$y^*_{i,t} = \boldsymbol{x}'_{i,t}\boldsymbol{\beta} + u_{i,t} + \alpha_i \tag{2}$$

with $E(U) = E(A) = 0$[7], and

$$y_{i,t} = \begin{cases} 1 & \text{if } y^*_{i,t} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Hereafter, I'll refer to the two models above, by assuming that either y^* (continuous response model) or y (binary response model) is observable. Estimates for the effect $\hat{\boldsymbol{\beta}}$ of the explanatory variables $\boldsymbol{x}$, net of the effects of UH, are obtained. If the model is linear, estimates of the individual intercepts $\hat{\alpha}_i$ are also obtained, together with estimates of the standard deviation of U and A. If the model is binary, a consistent estimate of the standard deviation of A is obtained with random effects, while only asymptotically biased estimates of the individual intercepts $\hat{\alpha}_i$ can be recovered with fixed effects.

The model must then be applied to a population $j = 1 \cdots N$, for which we know, at the beginning of the simulation at time $s = 0$, only the observable characteristics $\boldsymbol{x}_{j,0}$ and $y_{j,0}$. While $\hat{\boldsymbol{\beta}}$ can be directly used to construct the predicted outcome, a problem arises in assigning each simulated individual j a specific random intercept $\tilde{\alpha}_j$.

A simple solution is to set $\tilde{\alpha}_j = 0$ to each individual in the simulation sample. Another solution, which is possible only when estimates of the individual intercepts are available, is to treat the individual intercepts as missing variables which can be imputed

[7] I denote random variables with capital letters, and their realization with small letters.

by standard regression-based or matching techniques (Pickles, 2005; Howell, 2008) from their estimated value in the estimation sample. I'll now discuss what are the implications of these two strategies.

3 What if unobserved heterogeneity is neglected in forecasting?

Suppose that forecasts are evaluated on the basis of the mean squared forecasting error, at some future time s:

$$\text{MSFE}(\boldsymbol{v}_s) = \sum_j (v_{j,s} - \tilde{v}_{j,s})^2 / N \tag{4}$$

where $v = \{y^*, y\}$. We want to minimize the expected value of MSFE($\boldsymbol{v}$), and in particular we want to know whether (and possibly under which circumstances) setting $\tilde{\alpha}_j = 0 \; \forall j$ is a good forecasting strategy.[8]

As it turns out, the answer depends on whether we are considering a linear or a non-linear model. In linear models, there is a trade-off as imputing UH allows for a better description of the simulated individuals but at the same time introduces an additional noise factor in the forecasts: the key question therefore is whether $\tilde{A}$ is a decent predictor of A or not. In non-linear models, in addition to this trade-off, there is a non-negligible bias that is introduced in the projections if UH is not considered; this offers a strong case for imputing UH.

Proposition 1. *In the linear model of eq.* (2)*, setting* $\tilde{\alpha} = 0$ *leads to a higher MSFE unless* $\tilde{A}$ *is a poor predictor of* A*. If* $\hat{\boldsymbol{\beta}}$ *is consistent, or if the individual effects are uncorrelated with the observables, the condition for this to happen is* $\sigma_{A\tilde{A}} < \frac{\sigma^2_{\tilde{A}}}{2}$*, where* $\sigma_{A\tilde{A}}$ *is the covariance between the true and the imputed individual effects.*

[8] What I'm discussing here is setting $\tilde{\alpha}$ to 0 but using the UH-corrected coefficient $\hat{\boldsymbol{\beta}}$, not estimating the model without UH as for instance Panis (2003) does. Given that neglecting UH in forecasting implies an underestimation of the persistence of the outcome over time, one could consider estimating a model with lagged endogenous variable and without UH, and then use the estimated coefficients for forecasting. By doing so, the coefficient of the lagged endogenous variable (measuring true state dependency) would be biased upward, but will catch part of the effect of UH. Such specification would however fail to pass the Lucas critique and possibly lead to distorted policy prescriptions.

Proof. I compute for each individual the expected squared forecasting error (ESFE), and then I average over all individuals.

The expected squared forecasting error, given α_j and $\tilde{\alpha}_j$, is

$$\begin{aligned} \text{ESFE}_{j,s} &= E_U\left[(Y^*_{j,s} - \tilde{Y}^*_{j,s})^2 | \alpha_j, \tilde{\alpha}_j\right] \\ &= E_U\left[(\boldsymbol{x}'_{j,s}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + U + \alpha_j - \tilde{\alpha}_j)^2\right] \\ &= \sigma^2_U + (\alpha_j - \tilde{\alpha}_j)^2 + 2\boldsymbol{x}'_{j,s}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\alpha_j - \tilde{\alpha}_j) \end{aligned} \tag{5}$$

If $\hat{\boldsymbol{\beta}}$ is consistent, the last term vanishes as the sample size increases, and the condition for $\tilde{\alpha}_j = 0$ leading to a better forecast is $(\alpha_j - \tilde{\alpha}_j)^2 > \alpha_j^2$, which implies $\tilde{\alpha}_j(\tilde{\alpha}_j - 2\alpha_j) > 0$. This is satisfied, when α_j is positive, only if $\tilde{\alpha}_j < 0$ or $\tilde{\alpha}_j > 2\alpha_j$, and when α_j is negative, only if $\tilde{\alpha}_j < 2\alpha_j$ or $\tilde{\alpha}_j > 0$, that is when the imputed random intercept is far apart its true value.

On average over the simulated population, the expected value of the MSFE is

$$E\big[\text{MSFE}_s\big] = E_{A,\tilde{A}}\big[ESFE_{j,s}\big] = \sigma^2_U + \sigma^2_A + \sigma^2_{\tilde{A}} - 2\sigma_{A\tilde{A}} + 2(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})E_{\boldsymbol{x},A,\tilde{A}}\big[\boldsymbol{x}'_{j,s}(\alpha_j - \tilde{\alpha}_j)\big] \tag{6}$$

where $\sigma^2_{\tilde{A}} \to \hat{\sigma}^2_A$ is the estimated variance of the random intercept, which converges as the sample size grows bigger to the true variance, and $\sigma_{A\tilde{A}}$ is at most equal to σ^2_A (when the random intercepts are perfectly imputed).

The last term drops out if the individual effects are uncorrelated with the observables (as in a fixed effect model) or, in the limit, if the $\hat{\boldsymbol{\beta}}$ coefficients are consistent. The condition under which it becomes optimal not to impute UH is therefore

$$\sigma_{A\tilde{A}} < \frac{\sigma^2_{\tilde{A}}}{2} \tag{7}$$

that is, when $\tilde{A}$ is a poor predictor of A. □

In addition to the proposition above, it should be considered that neglecting UH (in simulation) introduces a break at the individual level between the past (for which outcomes depend on the true individual effect) and the future (for which outcomes depend

only on observables), thus making overall simulated life trajectories more cumbersome. This can have important consequences, for instance with respect to eligibility to social benefits, seniority accrual, *etc.*

The same issues arise in the non-linear case (although the math is more involved —see Appendix A). Moreover, in non-linear models there is an additional drawback from imputing a null UH to the simulated population, which shows up also when the simulated individuals enter the simulation without a history of previous outcomes. To see where it comes from, suppose two individuals have the same observables, but they differ because of UH: for the sake of illustration, suppose they have two symmetric individual effects around the mean value of 0. The outcome variable is binary. Non-linearity of the probit/logit transformation implies that the average probability of the event of the two heterogeneous individuals is in general different from the probability of the average individual, with the size of the bias depending on the standard deviation of the individual effect and the direction of the bias depending on the local concavity of the probit or logit transformation: imputing a null individual effect leads to overestimate the probability of the event if the average probability is higher than .5, and to underestimate it if it is lower.

Proposition 2. *In the binary response model of eq.* (3)*, setting* $\tilde{\alpha} = 0$ *leads to a bias in forecasting, irrespective of the previous history of the simulated individuals.*

Proof. In this setting, we have $\Pr\left[y_{j,s} = 1\right] = F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j\right)$ and $\Pr\left[\tilde{y}_{j,s} = 1\right] = F_U\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}} + \tilde{\alpha}_j\right)$. Let $z = \boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j$ and $\tilde{z} = \boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}} + \tilde{\alpha}_j$. A second order Taylor expansion of $F_U(z)$ and $F_U(\tilde{z})$ around $\boldsymbol{x}'_{j,s}\boldsymbol{\beta}$ and $\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}$ respectively gives

$$F_U(z) = F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta}\right) + f_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta}\right)\alpha_j + \frac{1}{2}f'_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta}\right)\alpha_j^2 + h_2\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta}\right)\alpha_j^2 \tag{8a}$$

$$F_U(\tilde{z}) = F_U\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}\right) + f_U\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}\right)\tilde{\alpha}_j + \frac{1}{2}f'_U\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}\right)\tilde{\alpha}_j^2 + h_2\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}\right)\tilde{\alpha}_j^2 \tag{8b}$$

and

$$E_A\left[F_U(z)|\boldsymbol{x}_{j,s}\right] = F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta}\right) + \frac{1}{2}f'_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta}\right)\sigma_A^2 \tag{9a}$$

$$E_{\tilde{A}}\left[F_U(\tilde{z})|\boldsymbol{x}_{j,s}\right] = F_Y\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}\right) + \frac{1}{2}f'_U\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}\right)\hat{\sigma}^2_A. \tag{9b}$$

Provided $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2_A$ are consistent, $E_{\tilde{A}}\left[F_U(\tilde{z})|\boldsymbol{x}_{j,s}\right]$ is a consistent estimator of $E_A\left[F_U(z)|\boldsymbol{x}_{j,s}\right]$: on average, the projections are correct, at the individual level. On the other hand, disregarding UH leads to $E_{\tilde{A}}\left[F_U(\tilde{z})|\boldsymbol{x}_{j,s}\right] = F_U\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}\right)$. The sign of the bias, $\frac{1}{2}f'_U\left(\boldsymbol{x}'_{j,s}\hat{\boldsymbol{\beta}}\right)\hat{\sigma}^2_A$, depends on the concavity of the CDF F_U, while its size depends on the variance of the individual effect.[9] □

In empirical applications the size of the bias is found to be of meaningful size. In a related paper, Richiardi and Poggi (2014) estimate a dynamic random effect microsimulation model of labor supply and household formation with lagged endogenous variables and compare the projections under three scenarios: no imputation of UH, imputation from the unconditional distribution of UH, and imputation from the conditional distributions of UH.[10] Using the standard deviation of the random effect of female labor supply estimated in that paper, it turns out that disregarding UH would artificially increase the participation rate in the relevant population from 57.1% to 61.0% in the base year, and from 65.0% to 73.2% in the final year of the simulation. Sampling from the unconditional estimated distributions of the individual effects prevents the forecasting bias, hence getting cross-sectional statistics right, but introduces unnatural breaks in individual trajectories at the moment of imputation, therefore getting longitudinal statistics wrong.[11]

After having motivated the need for imputing UH in the simulated sample, I now turn to the intuitive solution of treating the individual intercepts as missing variables which can be imputed by standard regression-based or matching techniques (Pickles, 2005; Howell, 2008) from their estimated value in the estimation sample, if available.

[9] Non-linearity of F also implies that the ML estimator for the $\boldsymbol{\beta}$ coefficients is biased.

[10] As a benchmark, a probit specification (without random effects) is also considered, thus addressing the issue raised in footnote 8.

[11] The same applies in the simple probit specification without UH.

4 Using the estimation sample as a donor dataset

With fixed effects, estimates of the individual intercepts $\hat{\alpha}_i$ can be obtained.[12] It may be therefore tempting to use the estimation sample as a donor dataset in order to impute the individual intercepts in the simulation sample. One simple way of doing this is to estimate

$$\hat{\alpha}_i = \boldsymbol{x}'_{i,t}\boldsymbol{\gamma} + \delta\upsilon_{i,t} + \varepsilon_{i,t} \tag{10}$$

where $\upsilon = \{y^*, y\}$, and then use the estimated coefficients $\hat{\boldsymbol{\gamma}}$ and $\hat{\delta}$ in the simulation sample to predict $\tilde{\alpha}_j$. Unfortunately, this approach has the drawback of distorting the distribution of the imputed individual intercepts.

Proposition 3. *In both the linear and the binary case, the distribution of the individual random intercepts in the simulation sample imputed using*

$$\tilde{\alpha}_j = \boldsymbol{x}'_{j,0}\hat{\boldsymbol{\gamma}} + \hat{\upsilon}_{j,0} \tag{10'}$$

where $\hat{\boldsymbol{\gamma}}$ and $\hat{\delta}$ are the coefficients of eq. (10) *estimated on the estimation sample, is in general different from the distribution of A.*

Proof. In the linear case, by substituting y^* in eq. (10′), we obtain

$$\tilde{\alpha}_j = \boldsymbol{x}'_{j,0}(\hat{\boldsymbol{\gamma}} + \hat{\delta}\hat{\boldsymbol{\beta}}) + \hat{\delta}(\alpha_j + u_j) \tag{11}$$

from which it is immediate to see that the distribution of $\tilde{A}$ is in general different from the distribution of A.[13]

[12] As I have already noted, such estimates are consistent in a linear setting, and inconsistent (due to the incidental parameters problem) in a binary setting.

[13] In particular, if the random effects assumption holds and the random disturbances A and U are normally distributed, $\hat{\boldsymbol{\gamma}} + \hat{\delta}\hat{\boldsymbol{\beta}} \to 0$ as the sample size increases, and the distribution of $\tilde{A}$ is still asymptotically normal. We have $\mathrm{var}(\tilde{A}) = \mathrm{var}(\tilde{A}|\boldsymbol{x}) = \delta^2\mathrm{var}(y^*|\boldsymbol{x}) \to \delta^2(\sigma_A^2 + \sigma_U^2)$ and $\hat{\delta} = \frac{\mathtt{cov}(\hat{A},y^*)}{\mathtt{var}(y^*)} = \frac{\sigma_A^2}{\sigma_A^2+\sigma_U^2}$, from which we obtain, as the sample size grows larger,

$$\mathrm{var}(\tilde{A}) \to \frac{\sigma_A^4}{\sigma_A^2 + \sigma_U^2} \neq \sigma_A^2$$

Similarly, in the binary case we get

$$\tilde{\alpha}_j = \boldsymbol{x}'_{j,0}\hat{\boldsymbol{\gamma}} + \hat{\delta}[\mathbf{1}(\boldsymbol{x}'_{i,t}\boldsymbol{\beta} + u_{i,t} + \alpha_i > 0)] \tag{12}$$

which is also not distributed as A. □

This regression-based imputation procedure matches individuals in the simulation sample with extrapolated individuals in the estimation sample, conditional on the specific values of the observables in the simulation sample. Such a conditioning however implies that in principle there might be no real individual, or more than one individual, in the estimation sample that can be used as a donor for the missing variable in the simulated individual. As a consequence, one has to resort to predicted values, and the uncertainty of the prediction is added to the aleatory distribution of the estimated residuals. As an alternative, it is possible to match each individual in the recipient dataset with a real individual in the donor dataset —at the cost of obtaining a less perfect match— in order to attribute the same estimated values of the random intercepts to the individuals in the simulation sample. This would guarantee that the distribution of $\tilde{A}$ resembles that of A. In the linear case, one natural propensity score matching would entail computing the estimated residuals $\hat{e}_i = \bar{y}^*_i - \bar{\boldsymbol{x}}'_i\hat{\boldsymbol{\beta}}$ and $\tilde{e}_j = y^*_{j,0} - \boldsymbol{x}'_{j,0}\hat{\boldsymbol{\beta}}$ respectively on the estimation and the simulation sample, and then match individuals based on these two variables ($\bar{y}_i$ and $\bar{\boldsymbol{x}}_i$ being the average values of the observed outcome and explanatory variables for individual i in the estimation sample). This is very similar to the imputation method I propose, with the only difference that rather than matching the estimated residuals in the simulation sample with their counterparts in the estimation sample, I match them with simulated residuals, built by sampling from the unconditional estimated distributions of A and U. This has the advantage that it only requires to estimate parametric distributions, rather than individual intercepts, and it is thus in principle amenable to application to settings where it is not possible to estimate individual intercepts —as in binary response random effects models— or not advisable to use them —for instance when the number of individuals in the estimation sample is too low.

This shift in focus from imputation of random intercepts estimated in the estimation sample to sampling from estimated distributions is common to other approaches in the literature, which I now turn to.

5 Sampling from the conditional distributions

According to the Bayesian approach, the correct way of assigning random intercepts $\tilde{\alpha}_j$ to individuals with a previous history of outcomes B is sampling from the estimated distribution of A, conditional on B. Applying Bayes theorem:

$$f_{A|B}(\alpha) = \frac{f_A(A=\alpha)\Pr(B|A)}{\Pr(B)} = \frac{f_A(A=\alpha)\Pr_U(B)}{\Pr_E(B)} \tag{13}$$

where $f_{A|B}$ and f_A are respectively the conditional and the unconditional distribution of A with respect to B, and $\Pr_U(B)$ and $\Pr_E(B)$ the conditional and unconditional distribution of B with respect to α.

Sampling from this distribution can be done by the Inverse Transform method: a random number is extracted from the uniform distribution on $[0,1]$, which gives the value of the conditional cumulative distribution function $F_{A|B}(\alpha) = \int f_{A|B}(\alpha)d\alpha$; $\tilde{\alpha}$ is then assigned by inverting this function, using the estimated distribution of A, $f_{\tilde{A}|B}$:

$$\begin{aligned} r &\sim U(0,1) \\ \tilde{\alpha} &= F_{\tilde{A}|B}(r)^{-1} \end{aligned} \tag{14}$$

In our continuous response linear model the conditional PDF is:

$$f_{A|Y^*=y^*}(\alpha, y^*) = \frac{f_A(A=\alpha)f_U(\boldsymbol{x'\beta} + U + \alpha = y^*)}{f_E(\boldsymbol{x'\beta} + E = y^*)} = \frac{f_A(\alpha)f_U(y^* - \boldsymbol{x'\beta} - \alpha)}{f_E(y^* - \boldsymbol{x'\beta})}. \tag{15}$$

If A and U are both Gaussian, with $A \sim N(0, \sigma_A^2)$ and $U \sim N(0, \sigma_U^2)$, this expression reduces to

$$f_{A|Y^*=y^*}(\alpha, y^*) \sim N\left(\frac{\xi_U(y^* - \boldsymbol{x'\beta})}{\xi_A + \xi_U}, \frac{1}{\xi_A + \xi_U}\right), \tag{16}$$

with $\xi_A = 1/\sigma_A^2$ and $\xi_U = 1/\sigma_U^2$. Individual effects are still normally distributed; when outcome is below expected outcome based on observables ($y^* - \boldsymbol{x'\beta} < 0$) they are centered on a negative number, and vice versa.

In the binary response model the conditional PDFs are:

$$f_{A|Y=0}(\alpha) = \frac{f_A(A=\alpha)\Pr(\boldsymbol{x'\beta} + U + \alpha < 0)}{\Pr(\boldsymbol{x'\beta} + U + A < 0)} = \frac{f_A(\alpha)(1 - F_U(\boldsymbol{x'\beta} + \alpha))}{1 - F_E(\boldsymbol{x'\beta})} \tag{17a}$$

$$f_{A|Y=1}(\alpha) = \frac{f_A(A=\alpha)\Pr(\boldsymbol{x'\beta} + U + \alpha > 0)}{\Pr(\boldsymbol{x'\beta} + U + A > 0)} = \frac{f_A(\alpha)F_U(\boldsymbol{x'\beta} + \alpha)}{F_E(\boldsymbol{x'\beta})} \tag{17b}$$

These conditional distributions do not have in general a closed form, even in the case when both the $\boldsymbol{\alpha}$ and the $\boldsymbol{u}$ are normal, though they can be approximated by a normal distribution with the Laplace method (Laplace, 1774, 1814). As an example, figure 1 depicts the conditional and unconditional distributions of the random intercepts, in a binary response model with $x \sim N(-0.5, 2), \beta = 1, A \sim N(0,1), U \sim N(0,1)$: it can be seen that the posterior distributions do indeed look quite normal.

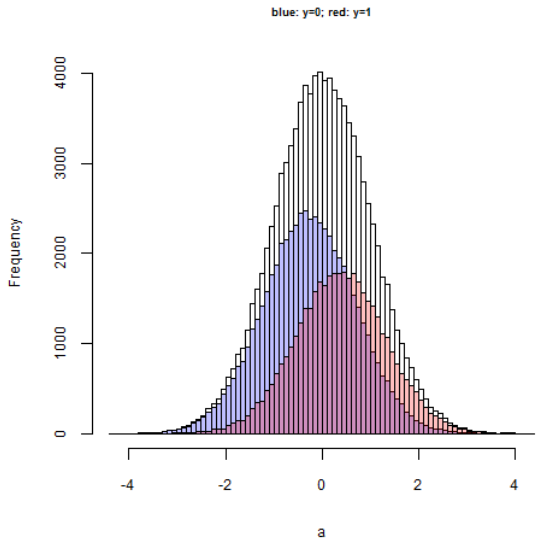


Figure 1: Conditional and unconditional distributions of random intercepts, binary response model. Parameterization: $x \sim N(-0.5, 2), \beta == 1, A \sim N(0,1), U \sim N(0,1), N = 100,000$.

Specifically, the posterior normal approximation is centered on the posterior mode, while the variance is estimated by looking at the curvature of the posterior at the maximum.[14] However, finding the mean and variance of the approximated normal distribu-

[14]Laplace Approximation methods are a family of deterministic algorithms that usually converges

tions, conditional both on y and $\boldsymbol{x}$, is not at all immediate, given that $\boldsymbol{x}$ might well be continuous.

Another approach is to reconstruct the conditional distributions by Markov Chain Montecarlo (MCMC) methods like Gibbs sampling (Casella and George, 1992; Bolstad, 2010). This is an algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution when direct sampling is difficult.[15] Unfortunately, application of this method is also not immediate, especially if the estimates have not been performed in a Bayesian framework —see Gu et al. (2009) for an application to a random effects binary probit model that allows for heteroscedasticity.

6 Optimal assignment

An alternative to the Bayesian procedure described above involves finding the distribution of individual effects that maximizes the (log)likelihood of observing the true data, $\ln L = \sum_{j=1}^{N} \ln L_{j,0}$. In the continuous response model, the individual contributions to the likelihood function are

$$L_{j,0}(\tilde{\alpha}) = \Pr(\tilde{Y}^* = y^*_{j,0} | A = \tilde{\alpha}_j) = f_U(y^*_{j,0} - \boldsymbol{x'}_{j,0}\hat{\boldsymbol{\beta}} - \tilde{\alpha}_j) \tag{18}$$

while in the binary response model:

$$L_{j,0}(0, \tilde{\alpha}) = \Pr(\tilde{Y} = 0 | A = \tilde{\alpha}_j) = 1 - F_U(\boldsymbol{x'}_{j,0}\hat{\boldsymbol{\beta}} + \tilde{\alpha}_j) \tag{19a}$$

$$L_{j,0}(1, \tilde{\alpha}) = \Pr(\tilde{Y} = 1 | A = \tilde{\alpha}_j) = F_U(\boldsymbol{x'}_{j,0}\hat{\boldsymbol{\beta}} + \tilde{\alpha}_j) \tag{19b}$$

Unconstrained maximization would require, in the linear case, setting $\tilde{\alpha}_j = y^*_{j,0} -$

faster than variational Bayes, much faster than MCMC, and just a little slower than Maximum Likelihood Estimation (MLE) (Azevedo-Filho and Shachter, 1994). However, the Laplace approximation shares many limitations of MLE, including asymptotic estimation with respect to sample size.

[15] In its basic version, Gibbs sampling is a special case of the Metropolis-Hastings algorithm. However, in its extended versions it can be considered a general framework for sampling from a large set of variables by sampling each variable (or in some cases, each group of variables) in turn, and can incorporate the Metropolis-Hastings algorithm (or similar methods such as slice sampling) to implement one or more of the sampling steps.

$\boldsymbol{x'}_{j,0}\hat{\boldsymbol{\beta}}$, that is, setting the individual intercept equal to the difference between the actual outcome $y^*_{j,0}$ and the predicted outcome based on observables only, $\boldsymbol{x'}_{j,0}\hat{\boldsymbol{\beta}}$. However, this would produce a distribution of the imputed random intercepts that resembles f_E, rather than f_A, and is not therefore consistent with the estimation results. The distortion is even clearer in the binary response case, where maximization of the likelihood is obtained respectively for $\tilde{\alpha}_j = \infty$ $(y_{j,0} = 1)$ and $\tilde{\alpha}_j = -\infty$ $(y_{j,0} = 0)$.

The solution therefore is a two-step procedure: first, N individual random intercepts are drawn from the estimated distribution of A; then, these random intercepts are optimally assigned to the N individuals in the simulation sample. Formally, we must find a permutation matrix $[p_{k,j}]$ where the index k refers to the random intercept and the index j to the individual to be simulated that solves (Koopmans and Beckmann, 1957)

$$\max_P \sum_{k,j} L_{k,j} p_{k,j} \tag{20}$$

$$\begin{aligned} &\text{s.t.} \\ p_{j,k} = \{0,1\} \quad & k, j = 1, \cdots, N \\ \textstyle\sum_j p_{k,j} = 1 \quad & k = 1, \cdots, N \\ \textstyle\sum_k p_{k,j} = 1 \quad & j = 1, \ldots, N \end{aligned}$$

The problem would in principle require to evaluate all $N!$ permutations of the individual effects: for $N = 100$, this number is 10×10^{157}. Standard optimal assignment algorithms as the Hungarian (or Kuhn-Munkres) algorithm reduce this problem to polynomial complexity (Carpaneto et al., 1988; Burkard et al., 2012). In particular, algorithms that are easier to implement have $O(N^4)$ complexity, while more complicated ones have $O(N^3)$ complexity. However, the assignment problem in a dynamic microsimulation model can in principle involve hundreds of thousands or even millions of individuals, which makes this approach feasible but still computationally burdensome.

An additional problem in the binary case is that any optimal assignment solution (more than one solutions generally exist, in a binary setting) introduces an artificial correlation between the imputed random intercepts and the observable characteristics.[16]

[16]The problem does not arise in the linear case, as the outcome fully reflects the value of the covariates.

To clarify, let's suppose that the true individual effects and the explanatory variables are indeed uncorrelated. The induced correlation problem arises from two sources that work in opposite directions. First, any optimal solution assigns, on average, smaller residuals to $y_0 = 0$ individuals, and bigger residuals to $y_0 = 1$ individuals. Since $y_0 = 0$ individuals have also, on average, lower values of the explanatory variables than $y_0 = 1$ individuals, this introduce a positive correlation between $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{x}$. Second, within each outcome group any optimnal solution assigns small residuals to individuals with high values of the explanatory variables (think of $y_0 = 0$ individuals with a high value of the score, which the algorithm tries to bring below zero) and big residuals to individuals with low values of the explanatory variables (think of $y_0 = 1$ individuals with a low value of the score, which the algorithm tries to bring above zero). Hence, within each outcome group a negative correlation between $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{x}$ is introduced.

Which effect prevails depend on the distribution of $\boldsymbol{x}$. If that distribution is narrow, so that the average value of the explanatory variables in the $y_0 = 0$ group is close to the average value of the explanatory variables in the $y_0 = 1$ group, the second effect predominates, and we get an overall negative correlation. If on the contrary the distribution of $\boldsymbol{x}$ is stretched, the first effect predominates, and we get an overall positive correlation.

To illustrate and test the relevance of the problem I set up to a Montecarlo analysis with the following parameterization: $N = 100, \sim N(0, \sigma_x), \beta = 1, A \sim N(0, 1), U \sim N(0, 1)$. For simplicity I assume again that β is estimated without errors: $\hat{\beta} = \beta$.

Figure 2 shows the value of the correlation coefficient between the imputed random intercepts $\hat{A}$ and the explanatory variable x, for the optimal solution of the assignment problem in the binary response model, for different values of σ_X. For low values of σ_X (UH is very important in determining the outcome) the correlation between $\hat{A}$ and X is negative; for higher values of σ_X the correlation turns positive.

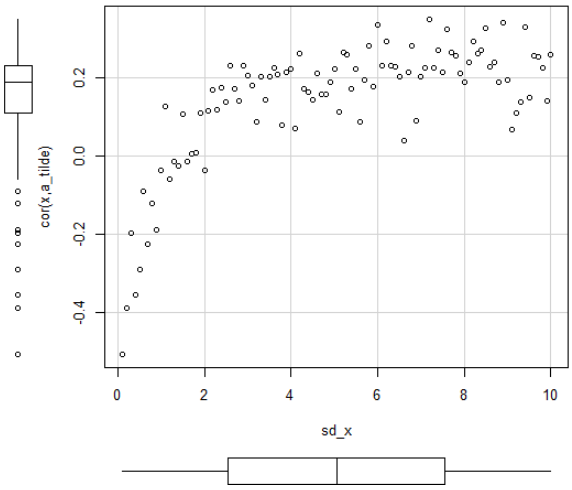


Figure 2: Correlation coefficient between the imputed random intercept $\tilde{A}$ and the explanatory variable x, optimal assignment, binary response model. Parameterization: $x \sim N(0, \sigma_x), \beta = \hat{\beta} = 1, A \sim N(0,1), U \sim N(0,1)$. For each value of σ_X 100 individuals are simulated. The Conditional Rank method described in appendix C is applied).

7 The Rank method

The method proposed here for solving the optimal assignment problem in linear models introduces two innovations on the approach reviewed above. First, it aims at assigning total residuals $\tilde{\boldsymbol{e}}_0$, rather than individual specific effects $\tilde{\boldsymbol{\alpha}}$.[17] The $\tilde{\boldsymbol{e}}_0$ are generated by summing up random draws from the unconditional $\hat{A}$ and $\hat{U}$ distributions, and their $\tilde{\boldsymbol{\alpha}}$ components are used as estimates of the true random intercepts $\boldsymbol{\alpha}$. Second, it relies on a new algorithm for assigning a $\tilde{e}_{j,0}$ to each individual in the simulated population. This algorithm is simpler to implement than the Hungarian method —hence it reduces programming time[18]— and involves a lower computational complexity —hence it reduces computing time.

The method works by minimizing the distance between the true errors $\boldsymbol{e}_0$ and the simulated errors $\tilde{\boldsymbol{e}}_0$, for the first (and only) period of observation in the simulation sample.

[17] In the linear case, setting $\tilde{e}_{j,0} = y^*_{j,0} - \boldsymbol{x'}_{j,0}\hat{\boldsymbol{\beta}}$ would maximize the likelihood, as we have already noted. An alternative solution of the assignment problem would require splitting each $\tilde{e}_{j,0}$ in its two components $\tilde{\alpha}_j$ and $\tilde{u}_{j,0}$, under the distributional constraints $f_{\tilde{A}} = f_{\hat{A}}$ and $f_{\tilde{U}} = f_{\hat{U}}$. Although conceptually very simple, this approach is computationally quite demanding —except for very specific assumptions about the two distributions— and requires in general multiple loops over the simulated population, each involving extractions from the unconditional $\hat{A}$ and $\hat{U}$ distributions, until the distributional constraints are satisfied.

[18] The method requires no more than a handful lines of code, while standard Matlab or Java implementations of the Hungarian algorithm consist in about 500 lines of code.

However, the true errors are unobservables. Therefore, the method minimizes the distance between some proxy of the true errors $\boldsymbol{e}_0$ and the simulated errors $\tilde{\boldsymbol{e}}_0$. As $\tilde{\boldsymbol{e}}_0$ are given, the task is an assignment problem.

An obvious choice for this proxy in linear models are the regression residuals, that is the difference between the observed outcome and the outcome predicted on the observables only.

Specifically, the Rank method works as follows:

Algorithm 1. ***Rank method***

1. *Estimate the random intercept model (on the estimation sample).*
2. *Compute (on the simulation sample) the predicted outcome* $\hat{\boldsymbol{y}}_0^* = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ *by imposing* $\tilde{\alpha}_j = 0 \ \forall j$.
3. *Compute the difference between the observed outcome and the predicted outcome based on observables only,* $\boldsymbol{y}_0 - \hat{\boldsymbol{y}}_0$*, and order this difference from high to low (with randomized tie-breaking).*
4. *Extract* N *values from the (parametric or empirical) unconditional distribution of the individual intercept,* $\tilde{\boldsymbol{\alpha}}$.
5. *Extract* N *values from the (parametric or empirical) unconditional distribution of the idiosyncratic disturbance,* $\tilde{\boldsymbol{u}}_0$.
6. *Construct the error terms* $\tilde{\boldsymbol{e}}_0 = \tilde{\boldsymbol{\alpha}} + \tilde{\boldsymbol{u}}_0$ *and order them from high to low.*
7. *Assign the random intercepts* $\tilde{\boldsymbol{\alpha}}$ *to the individuals by matching the two rankings described above.*

Given that the Rank method only requires to match two rankings, it works in $N \log N$ (quasilinear) time, as the simple Montecarlo experiment reported in figure 3 shows.

7.1 An example

Here I provide a numerical example of the Rank method. Suppose that a random effects estimation of eq. (2) gives a vector of estimated coefficients $\hat{\boldsymbol{\beta}}$ and an estimate $\hat{\sigma}_A$ for

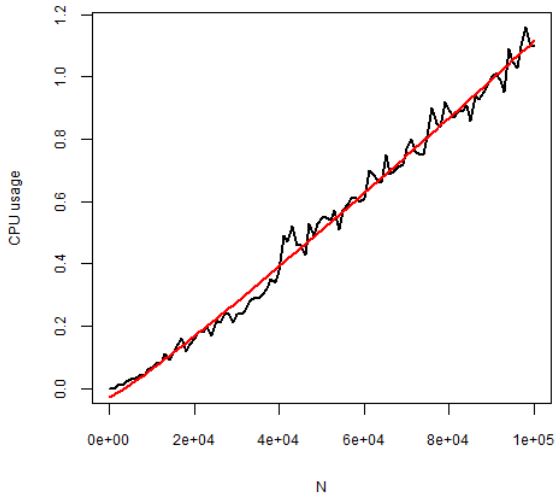


Figure 3: Speed of execution as a function of sample size, Rank method. Parameterization: $x \sim N(0,1), \beta = \hat{\beta} = 1, A \sim N(0,1), U \sim N(0,1)$. The $N \log N$ fit is superimposed (red line).

the standard deviation of the random effects. An estimate of the standard deviation of the error term $\hat{\sigma}_U$ is also obtained. I assume that $\hat{\boldsymbol{\beta}}$ are estimated without errors ($\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$). Random intercepts $\tilde{\boldsymbol{\alpha}}$ have to be assigned to individuals in the simulation sample. Columns (1)-(3) in table 1 reconstruct the outcome. Column (4) ranks the difference between outcome and predicted outcome, assuming no individual effects (because of the assumption that $\boldsymbol{\beta}$ is estimated without errors, this difference is equal to $\boldsymbol{e}_0$). All these values refer to the observed data. The last four columns refer to imputation. Column (5) reports the value of the imputed individual effects $\tilde{\boldsymbol{\alpha}}$, drawn from a normal distribution with mean 0 and standard deviation $\hat{\sigma_A}$. Column (6) contains random draws $\tilde{\boldsymbol{u}}_0$ from a normal distribution with mean 0 and standard deviation $\hat{\sigma_U}$. Columns (5) and (6) add to the total residuals $\tilde{\boldsymbol{e}}_0$, which are displayed in column (7) and ranked in column (8).

The Rank method assigns high draws of the total residuals, which on average are associated to high draws of the random intercepts, to individuals with a high value of the outcome variable and a low value of the score, and vice versa. For instance, the observable characteristics of individual *6* point to a low value of the outcome. The fact that this is not true suggests that she has some positive unobservable characteristics; accordingly, individual *6* is assigned a high draw of the total residual $\tilde{e}_0$. The stochastic nature of the algorithm allows such a draw to be associated to a low value of the imputed individual

Table 1: Application of the Rank method in a linear regression model.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
id	$\boldsymbol{x}_0\boldsymbol{\beta}$	e_0	y_0^*	$r(y_0^* - \boldsymbol{x}_0\hat{\boldsymbol{\beta}})$	$\tilde{\alpha}$	$\tilde{u}_0$	$\tilde{e}_0$	$r(\tilde{e}_0)$
6	-0.77	2.62	1.85	1	-1.43	2.01	0.58	1
7	-0.49	2.24	1.75	2	2.36	-2.12	0.24	2
5	-1.38	1.77	0.39	3	0.22	-0.10	0.12	3
8	-0.39	1.70	1.32	4	2.02	-2.45	-0.43	4
10	0.21	0.83	1.05	5	-0.76	0.16	-0.60	5
9	-0.05	0.32	0.28	6	-1.35	0.27	-1.08	6
3	-0.37	-0.47	-0.84	7	0.29	-1.45	-1.16	7
1	-2.79	-0.51	-3.30	8	-1.87	0.06	-1.81	8
2	-0.50	-0.64	-1.14	9	-0.52	-1.85	-2.37	9
4	0.36	-0.90	-0.54	10	0.79	-3.33	-2.54	10

Note: It is assumed $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

effect $\tilde{\alpha}$, although this happens with a low probability.

Table 2 describes the application of the Rank method to the binary case of eq. (3): a binary outcome y is now observed (column (3) in table 2), in lieu of y^*, following eq. (3). Column (4) displays the value of the predicted outcome, based on the observables only: $\hat{y}_0 = \mathbf{1}\big(\boldsymbol{x}'_{j,0}\hat{\boldsymbol{\beta}} > 0\big)$; column (5) contains the difference between observed and predicted outcome; column (6) shows the ranking of this difference, assuming random tie-breaking rule. Columns (7)-(10) are the same as in the right hand side of table 1. Finally, column (11) reports the value of the predicted outcome, once the imputed random intercept is taken into consideration: $\tilde{y}_0 = \mathbf{1}\big(\boldsymbol{x}'_{j,0}\hat{\boldsymbol{\beta}} + \tilde{e}_{j,0} > 0\big)$.

Table 2: Application of the Rank method in a binary response model.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
id	$\boldsymbol{x}_0\boldsymbol{\beta}$	e_0	y_0	$\hat{y}_0$	$y_0 - \hat{y}_0$	$r(y_0 - \hat{y}_0)$	$\tilde{\alpha}$	$\tilde{u}_0$	$\tilde{e}_0$	$r(\tilde{e}_0)$	$\tilde{y}_0$
7	-0.49	2.24	1	0	1	1	-1.43	2.01	0.58	1	1
9	-0.05	0.32	1	0	1	2	2.36	-2.12	0.24	2	1
8	-0.39	1.70	1	0	1	3	0.22	-0.10	0.12	3	0
5	-1.38	1.77	1	0	1	4	2.02	-2.45	-0.43	4	0
6	-0.77	2.62	1	0	1	5	-0.76	0.16	-0.60	5	0
10	0.21	0.83	1	1	0	6	-1.35	0.27	-1.08	6	0
1	-2.79	-0.51	0	0	0	7	0.29	-1.45	-1.16	7	0
2	-0.50	-0.64	0	0	0	8	-1.87	0.06	-1.81	8	0
3	-0.37	-0.47	0	0	0	9	-0.52	-1.85	-2.37	9	0
4	0.36	-0.90	0	1	-1	10	0.79	-3.33	-2.54	10	0

Note: It is assumed $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

The same logic as in the linear regression case applies, but now the nonlinearity of

the model hides much information. The MSFE is in this case simply the fraction of observations for which the predicted outcome $\tilde{y}_0 = \mathbf{1}\big(\boldsymbol{x}'_{j,0}\hat{\boldsymbol{\beta}} + \tilde{e}_{j,0} > 0\big)$ does not match the actual outcome $y_0 = \mathbf{1}(\boldsymbol{x}'_{j,0}\boldsymbol{\beta} + e_{j,0} > 0)$. In this example, it amounts to 4 (individuals *5*, *6*, *8* and *10*).

7.2 Properties of the Rank method, linear case

The following two conditions are sufficient for optimality of the Rank method, in the linear case: (i) minimizing the mean squared error $\text{MSE}(\hat{\boldsymbol{e}}_0) = \sum_j (e_{j,0} - \tilde{e}_{j,0})^2$ leads to optimal forecasts (for a given extraction of $\tilde{\boldsymbol{e}}_0$), (ii) matching the two rankings of $\hat{\boldsymbol{e}}_0$ and $\tilde{\boldsymbol{e}}_0$ minimizes $\text{MSE}(\hat{\boldsymbol{e}}_0)$ (for the given extraction of $\tilde{\boldsymbol{e}}_0$). I now establish that they both hold, as sample size increases.

Proof of condition (i) is trivial. From $\tilde{y}^*_{j,0} = \boldsymbol{x}'_{j,0}\hat{\boldsymbol{\beta}} + \tilde{e}_{j,0}$ we immediately get

$$y^*_{j,0} - \tilde{y}^*_{j,0} = \boldsymbol{x}'_{j,0}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \hat{e}_{j,0} - \tilde{e}_{j,0}. \tag{21}$$

Hence, if $\hat{\boldsymbol{\beta}}$ is consistent, minimizing $\text{MSE}(\hat{\boldsymbol{e}}_0)$ is equivalent to minimizing $\text{MSFE}(\boldsymbol{y}_0)$.

The following proposition states that the Rank method does indeed solve the optimal assignment problem in linear regression models:

Proposition 4. *Given the model* (2)*, the Rank solution obtained by applying algorithm 1 minimizes MSE*$(\hat{\boldsymbol{e}}_0) = \sum_j (\hat{e}_{j,0} - \tilde{e}_{j,0})^2/N$.

Proof. See appendix B. □

8 Properties of the Rank method applied to the binary response case

In the binary response case, the Rank method is not optimal. To see why, consider switching the imputed random intercepts for individuals *8* and *10*: while individual *8*, who is observed as a success, remains predicted as a failure, individual *10* (who is also

a success) is now assigned a positive value of UH, and therefore turns concordant. Said differently, assigning a positive but small value of the random intercept to *8* is a waste, as it does not make the score of *8* positive.[19] However, it turns out that applying the Rank method to a binary setting, while leading to a decrease in the ability to replicate the observed data with respect to the optimal assignment solution, obtains in exchange a drastic reduction in the artificial correlation that is introduced between the imputed individual effects and the explanatory variables. Indeed, given that other simple methods of imputing the individual effects are available in the linear case, the biggest value added of the Rank method lies in its application to the nonlinear case.

The left panel of figure 4 refers to the Rank method in the linear regression model, and shows that no correlation is found which was not present in the data generating process. The right panel shows the performance of the Rank method in the binary response model. The values of the parameters are the same as in the Montecarlo experiment of figure 2.

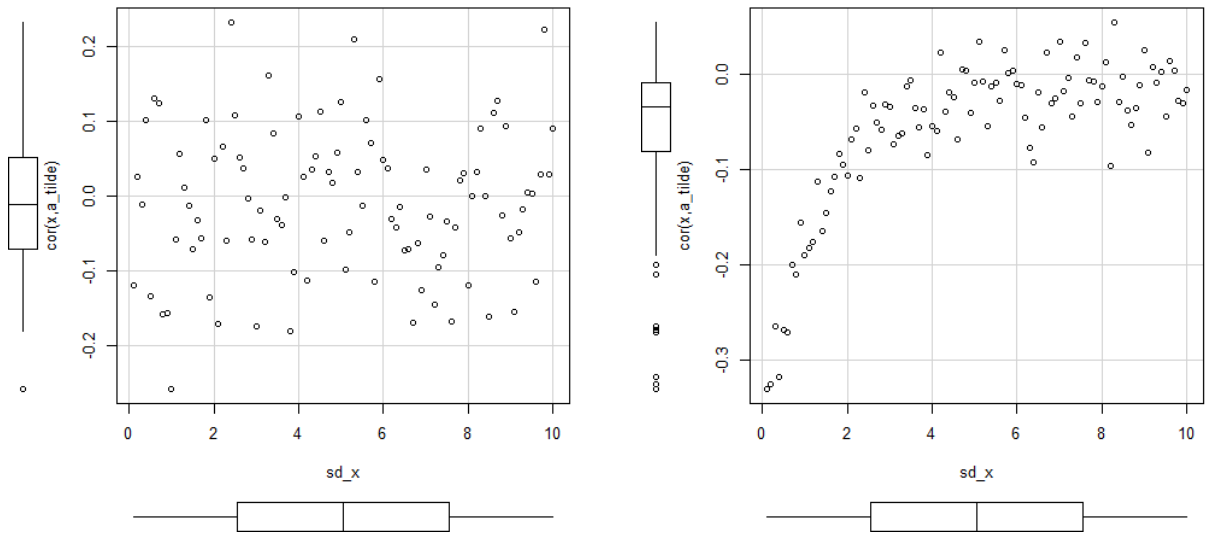


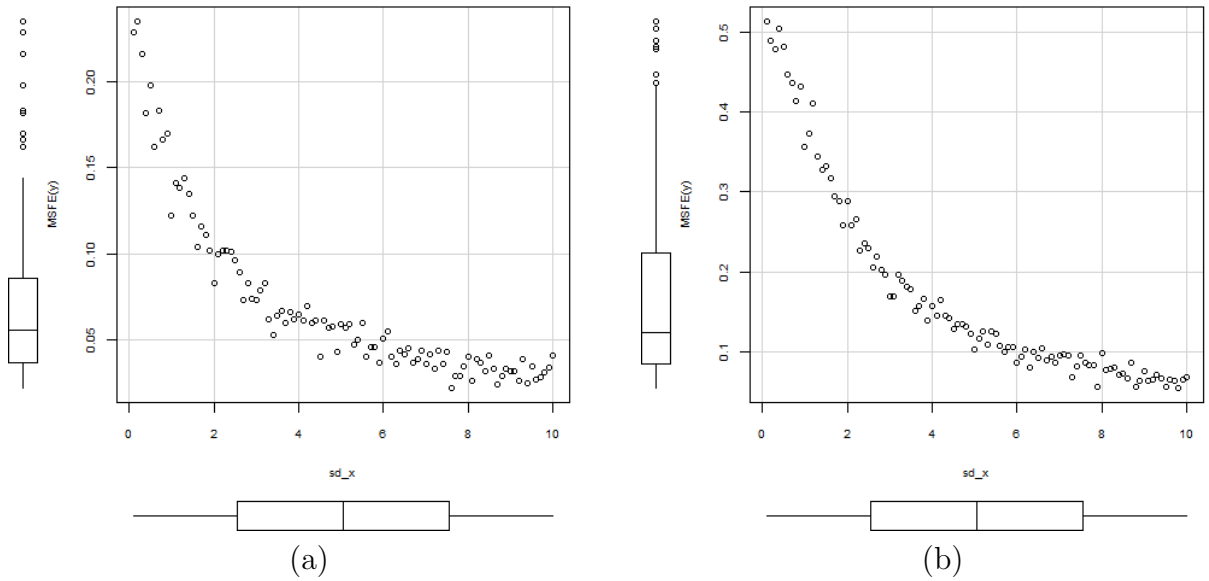
Figure 4: Correlation coefficient between the imputed random intercept $\tilde{A}$ and the explanatory variable x, Rank method. Parameterization: $x \sim N(0, \sigma_x), \beta = \hat{\beta} = 1, A \sim N(0,1), U \sim N(0,1)$. For each value of σ_X 100 individuals are simulated and the Rank method is applied. Left panel (a): continuous response model. Right panel (b): binary response model.

Only when the explanatory variables are relatively unimportant in determining the

[19] In appendix C I develop a variation of the Rank method that solves the optimal assignment problem in the binary response case in quadratic rather than cubic time, as the standard Hungarian method. The algorithm also clarifies the mechanics behind the induced correlation problem.

outcome the correlation is significant. Its sign depends on the local curvature of the logit/probit transformation around the mean value $\bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}$, as individuals are ranked according to $\boldsymbol{y} - \hat{\boldsymbol{y}} = \mathbf{1}(\boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{e} > 0) - \mathbf{1}(\boldsymbol{x}\hat{\boldsymbol{\beta}} > 0)$. As the variance of $\boldsymbol{x}$ becomes bigger, the induced correlation becomes very small and then vanishes. This corresponds to the fact that, due to the discrete nature of the outcome, with very large or very low values of the score UH makes little difference. Therefore, $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ tend to be concordant, and the random intercepts are assigned randomly.

Figure 5 shows the performance of the Rank method in a binary response setting (left panel), and compares it with a pure random assignment (right panel). As UH looses importance (σ_X increases with respect to σ_A), the Rank method converges to a random assignment. However, the Rank method still roughly halves the MSFE, with respect to random assignment (as a further benchmark, consider that the Conditional Rank method, which solves the optimal assignment problem in binary response models, drives the MSFE basically to zero).



(a) (b)

Figure 5: Mean squared forecasting error $\text{MSFE}(\boldsymbol{y}_0) = \sum_{j=1}^{N}(y_{j,0} - \tilde{y}_{j,0})$, binary response model. Parameterization: $x \sim N(0, \sigma_x), \beta = \hat{\beta} = 1, A \sim N(0,1), U \sim N(0,1)$. For each value of σ_X 100 individuals are simulated. Left panel (a): Rank method. Right panel (b): random assignment.

Even when the random disturbances play a little role in determining the outcome, for high values of σ_x^2, the MSFE with the Rank method is still about half the size than with

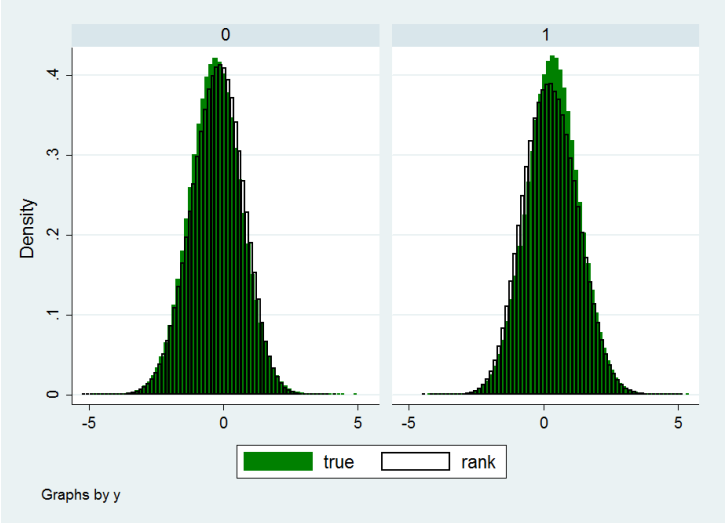


Figure 6: Distribution of the true and imputed (via the Rank method) individual effects. Parameterization: $x \sim N(-0.5, 2), \beta = \hat{\beta} = 1, A \sim N(0,1), U \sim N(0,1)$

unconditional assignment (for comparison, the optimal assignment solution brings the MSFE down to practically zero irrespective of σ_x^2). Finally, the Rank method is also able to replicate to a satisfactory extent the conditional distribution properties of the true individual effects. Figure **??** shows the distributions of the true and imputed individual effects for the $y = 0$ and $y = 1$ subsamples, in an additional Monte Carlo experiment where 1 million observations were drawn according to $x \sim N(-0.5, 2), \beta = \hat{\beta} = 1, A \sim N(0,1), U \sim N(0,1)$.

For extreme values of x, the Rank method introduces some distortions in the distribution of the individual effects, but still much lower than random assignment (that is, sampling from the unconditional distribution of the individual effects). This can be seen in figure 7, for the first quintile of x.

9 Summary and conclusions

In this paper I have dealt with the problem of forecasting the evolution of a population where the estimated processes account for unobserved heterogeneity. If the simulation sample does not coincide with the estimation sample and the initial conditions of the simulation include information on outcomes, the problem is non-trivial. This situation is indeed quite common, for instance in dynamic microsimulation modeling, and involves

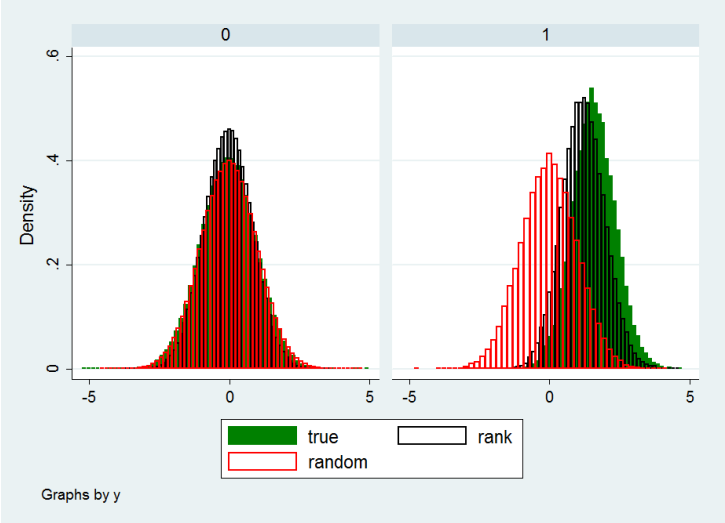


Figure 7: Distribution of the true and imputed (via the Rank method) individual effects, first quintile of x. Parameterization: $x \sim N(-0.5, 2), \beta = \hat{\beta} = 1, A \sim N(0, 1), U \sim N(0, 1)$.

assigning individual effects to the simulated population by sampling from the conditional estimated distributions of the individual effects (conditional on the observed outcomes), rather than from the unconditional distributions. This is straightforward in linear models but involves the use of MCMC methods in nonlinear models.

The solution I have proposed in this paper is to draw from the unconditional distribution of the overall error term, composed by an individual intercept plus a random component, and then assign the individual-specific component of this overall error term to the agents in the simulation. I have presented an algorithm which is computationally very efficient and solves the optimal assignment problem in linear settings; applied to nonlinear models, it offers a good approximation of the distributional properties of the individual effects. The intuition behind the algorithm is using the difference between observed and predicted outcome as a proxy of the unobserved individual component, which is then matched with the sampled residuals. The algorithm can be thought of as variants of propensity score matching, where individuals in the simulation sample are matched not with individuals in the estimation sample, but with simulated counterparts sampled from estimated unconditional distributions.

For nonlinear applications where the approximation offered by the Rank method is not enough, and the use of MCMC methods like Gibbs sampling too costly, the use of a linear probability model should be considered. The common solution of neglecting unobserved

heterogeneity in forecasting is shown to badly affect the quality of the forecasts, especially in nonlinear models.

A Neglecting UH in binary response models

In binary response models, the outcome collapses on the two distinct values 0 and 1. *Ex post*, the best option is to forget UH when the random draw for the imputed UH is negative while the realized outcome is 1 (as the negative draw decreases the probability of simulating a success, hence increasing the MSFE), and when the random draw for UH is positive, while the realized outcome is 0 (as the positive draw decreases the probability of simulating a failure, hence increasing again the MSFE). An *ex ante* decision rule would then require the evaluation of the probability of obtaining a success/failure for all different values of $\boldsymbol{x}$, A and $\tilde{A}$. The multiple integration required however proves impossible to deal with analytically.

Let's assume, as a first approximation, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.[20] The expected forecasting error, for given α_j, $\tilde{\alpha}_j$ and $\boldsymbol{x}_{j,s}$, is

$$
\begin{aligned}
\mathrm{ESE}_{j,s} =& E_U\left[y_{j,s} - \tilde{y}_{j,s}\right] \\
=& \Pr\left[y_{j,s} = 1 \wedge \tilde{y}_{j,s} = 0\right] + \Pr\left[y_{j,s} = 0 \wedge \tilde{y}_{j,s} = 1\right] \\
=& \Pr\left[y_{j,s} = 1\right] \Pr\left[\tilde{y}_{j,s} = 0 | y_{j,s} = 1\right] + \Pr\left[y_{j,s} = 0\right] \Pr\left[\tilde{y}_{j,s} = 1 | y_{j,s} = 0\right] \\
=& F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j\right) \Pr\left[\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \tilde{\alpha}_j + \tilde{u}_{j,s} < 0 | \boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j + u_{j,s} \geq 0\right] \\
&+ \left[1 - F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j\right)\right] \Pr\left[\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \tilde{\alpha}_j + \tilde{u}_{j,s} \geq 0 | \boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j + u_{j,s} < 0\right] \\
=& F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j\right) F_{U-\tilde{U}}\left(\alpha_j - \tilde{\alpha}_j + z_{j,s}(\boldsymbol{x})\right) \\
&+ \left[1 - F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j\right)\right]\left[1 - F_{U-\tilde{U}}\left(\alpha_j - \tilde{\alpha}_j + z_{j,s}(\boldsymbol{x})\right)\right] \\
=& 1 - F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j\right) - F_{U-\tilde{U}}\left(\alpha_j - \tilde{\alpha}_j + z_{j,s}(\boldsymbol{x})\right) + 2F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j\right) F_{U-\tilde{U}}\left(\alpha_j - \tilde{\alpha}_j + z_{j,s}(\boldsymbol{x})\right)
\end{aligned}
\tag{22}
$$

with $\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j + u_{j,s} + z_{j,s}(\boldsymbol{x}) = 0$.

The condition when disregarding $\tilde{\alpha}$ leads to a better forecast is

$$
F_{U-\tilde{U}}\left(\alpha_j - \tilde{\alpha}_j + z_{j,s}\right) - F_{U-\tilde{U}}\left(\alpha_j + z_{j,s}\right) < 2F_U\left(\boldsymbol{x}'_{j,s}\boldsymbol{\beta} + \alpha_j\right)\left[F_{U-\tilde{U}}\left(\alpha_j - \tilde{\alpha}_j + z_{j,s}\right) - F_{U-\tilde{U}}\left(\alpha_j + z_{j,s}\right)\right] \tag{23}
$$

[20] This is wrong even in expected terms, as the probit/logit coefficients (and, in general, non-linear estimators) are not unbiased.

which gives

$$\begin{cases} \Pr\left[y_{j,s} > .5\right] & \text{if } \tilde{\alpha}_j < 0 \\ \Pr\left[y_{j,s} < .5\right] & \text{if } \tilde{\alpha}_j > 0 \end{cases} . \tag{24}$$

This translates, after integration over $\boldsymbol{x}, A, \tilde{A}$, on a condition on the covariance between A and $\tilde{A}$, which is however much more complicated than in linear models (see Proposition 1).

B Proof of Proposition

Given the model (2)*, the Rank solution obtained by applying algorithm 1 minimizes* $MSE(\hat{\boldsymbol{e}}_0) = \sum_j (\hat{e}_{j,0} - \tilde{e}_{j,0})^2/N$.

The proof is organized in two parts. First, I show that the Rank solution is optimal for $N = 2$. Second, I show that when $N > 2$ a non-Rank solution can be improved upon by another non-Rank solution where non-Rank connections involving any 2 couples $(\hat{e}, \tilde{e})$ are replaced using the $N = 2$ result. This process can be repeated until the Rank solution is obtained.

For simplicity I drop the time index 0, which specifies that the estimated residuals $\hat{\boldsymbol{e}}$ depend on the idiosyncratic term $\boldsymbol{u}_0$ and that the simulated residuals $\tilde{\boldsymbol{e}}$ depend on the random extraction $\tilde{\boldsymbol{u}}_0$. Denote as $\Delta_j = \hat{e}_j - \tilde{e}_j$ the argument of the individual contribution to the objective function. Also, let $S = \sum_{j=1}^{N} |\Delta_j|$. In what follows numerical indexes for $\hat{e}$ and $\tilde{e}$ stand for their rank, and not to the individuals they refer to. Hence, $\hat{e}_1 \leq \hat{e}_2 \leq \cdots \leq \hat{e}_N$ and $\tilde{e}_1 \leq \tilde{e}_2 \leq \cdots \leq \tilde{e}_N$. The position in the ranking is generically identified with the letter k. As an example, $\tilde{e}_j = \tilde{e}_k$ means that individual j is assigned the kth biggest (or smallest, for what matters) value for her random intercept, among those extracted.

Lemma 1. *The Rank solution minimizes* $MSE(\hat{\boldsymbol{e}}_0)$ *when* $N = 2$.

Proof. For ease of visualization, I draw the $\hat{\boldsymbol{e}}_k$ and the $\tilde{\boldsymbol{e}}_k$ on two parallel axes. In an $N = 2$ case, the four points $\{\hat{e}_1, \hat{e}_2, \tilde{e}_1, \tilde{e}_2\}$ define a quadrilateral: the Rank solution

involves connecting the edges, while the (unique) non-Rank solution involves connecting the diagonals. Define a and d the two edges (connecting respectively $\hat{e}_1$ to $\tilde{e}_1$ and $\hat{e}_2$ to $\tilde{e}_2$), while b and c the two diagonals (connecting $\hat{e}_1$ to $\tilde{e}_2$ and $\hat{e}_2$ to $\tilde{e}_1$).

According to the ordering of the vertex, there are 6 possible cases (figure 8). Symmetry of cases (iii)-(iv) and (v)-(vi) allows us to focus on cases (i), (ii), (iii) and (v) only.

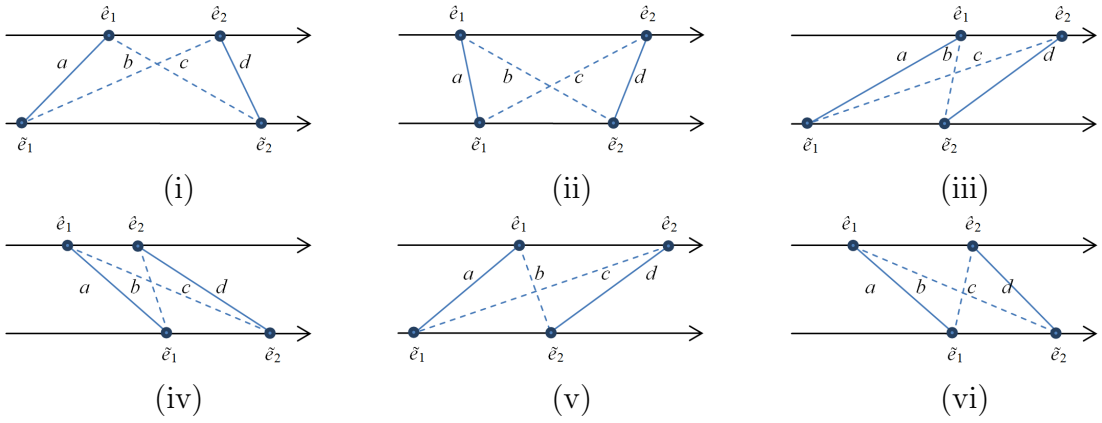


Figure 8: $N = 2$, different orderings of $\{\hat{e}_1, \hat{e}_2, \tilde{e}_1, \tilde{e}_2\}$.

Case (i): $\tilde{e}_1 \leq \hat{e}_1 \leq \hat{e}_2 \leq \tilde{e}_2$. It is immediate to see that the diagonals are longer than the edges: $a \leq c, d \leq b$. Hence, MSE($\hat{\boldsymbol{e}}$) is minimized with the Rank solution.

Case (ii): $\hat{e}_1 \leq \tilde{e}_1 \leq \tilde{e}_2 \leq \hat{e}2$. Again, the diagonals are longer than the edges: $a \leq b, d \leq c$.

Case (iii): $\tilde{e}_1 \leq \tilde{e}_2 \leq \hat{e}_1 \leq \hat{e}_2$. In this case, $S_{\text{R}} = S_{\text{NR}}$. In facts, $S_{\text{R}} = \hat{e}_1 - \tilde{e}_1 + \hat{e}_2 - \tilde{e}_2$, while $S_{\text{NR}} = \hat{e}_1 - \tilde{e}_2 + \hat{e}_2 - \tilde{e}_1$.[21] Now, concavity of the objective function implies that, for any given S, MSE($\hat{\boldsymbol{e}}$) is minimized when heterogeneity in Δ is kept at a minimum.[22] Simple arithmetic shows that, for fixed S, $\text{var}(\Delta)_{\text{R}} \leq \text{var}(\Delta)_{\text{NR}}$ implies $(\hat{e}_2 - \hat{e}_1)(\tilde{e}_2 - \tilde{e}_1) \geq 0$, which is true by construction. Therefore, the variance of Δ in the Rank solution is never greater than in the non-Rank solution, and the Rank solution is again optimal.

Case (v): $\tilde{e}_1 \leq \hat{e}_1 \leq \tilde{e}_2 \leq \hat{e}_2$. Keep fixed $\tilde{e}_1, \hat{e}_1$ and $\hat{e}_2$. The best case for the non-Rank solution, and the worse for the Rank solution, is $\tilde{e}_2 = \hat{e}_1$ (in which case $b = 0$ and d, given

[21] More in general, it is possible to show that the necessary conditions for $S_{\text{R}} = S_{\text{NR}}$ are:

$$\begin{aligned} &\hat{e}_k \geq \tilde{e}_{k+1} \quad \forall k \in \{0, N-1\} \quad \text{or} \\ &\hat{e}_{k+1} \leq \tilde{e}_k \quad \forall k \in \{0, N-1\} \end{aligned}$$

.

[22] This stems immediately from the definition of a concave function: $f(sm + (1-s)M) \geq sf(m) + (1-s)f(M)$.

the constraints, is maximum). This is a borderline case (iii) situation, and the result for case (iii) applies. *A fortiori*, the non-Rank solution cannot be optimal when $\tilde{e}_2 > \hat{e}_1$.

Similar arguments can be made for case (iv) —symmetric to case (iii)— and case (vi) —symmetric to case (v). Hence, the Rank solution is optimal when $N = 2$. □

Lemma 2. *Subsequent replacements of diagonals with edges when $N > 2$ improve the MSE and ultimately lead to the Rank solution.*

Proof. Take any two connections in a non-Rank solution in the $N > 2$ case, say those for individuals i and j, for which $\hat{e}_i < \hat{e}_j$ and $\tilde{e}_i > \tilde{e}_j$. Then, the result of lemma 1 applies and the non-Rank solution can be improved by connecting the smallest $\hat{e}$ with the smallest $\tilde{e}$, and the largest $\hat{e}$ with the largest $\tilde{e}$, that is rewiring according to the partial ranks. Now, only two cases are possible: either the new connections do not cross any other connection, that is $\nexists h$ such that $\hat{e}_h \lesseqgtr \hat{e}_i, \tilde{e}_h \gtreqless \tilde{e}_i$ or $\hat{e}_h \lesseqgtr \hat{e}_j, \tilde{e}_h \gtreqless \tilde{e}_j$, in which case we have obtained the Rank solution, or such an h exists, and the process of replacement of diagonals with edges can start again.

Figure 9 shows an example.

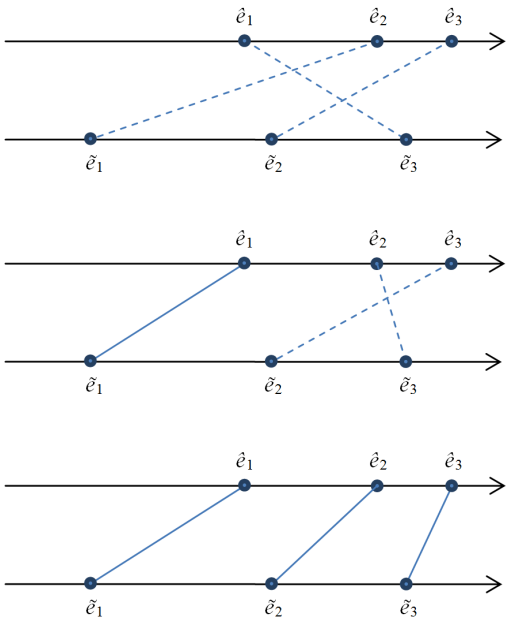


Figure 9: Subsequent replacement of diagonals with edges.

□

Together, the two lemmas above imply optimality of the Rank solution.

C The Conditional Rank method

To solve the optimal assignment problem in binary response settings, the Rank method has to undergo two (minor) modifications. First, to avoid throwing away information which is levelled off by the discrete nature of the outcome variable, in binary response models it is better to use $-\boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ —the difference between the threshold in the latent variable that determines outcome (which is normalized to 0) and the predicted value of the score— as a proxy of the true errors, rather than $y - \hat{y}$. Second, in order not to "waste" random terms that could be helpful in matching other observations by assigning them to individuals for which they don't make a difference, however, an additional control needs to be made, separately on successes ($y_{j,0} = 1$) and failures ($y_{j,0} = 0$).

Following closely this constructive approach, an algorithm for solving the optimal assignment problem in binary response models can be defined:

Algorithm 2. ***Conditional Rank method for binary response models***

1. *Estimate the random intercept model (on the estimation sample).*
2. *Compute the score* $\boldsymbol{x}_0\hat{\boldsymbol{\beta}}$*, by imposing* $\tilde{\alpha}_j = 0\ \forall j$ *(on the simulation sample), and rank the score from high to low.*
3. *Extract* N *values from the (parametric or empirical) unconditional distribution of the individual intercept,* $\tilde{\boldsymbol{\alpha}}$*.*
4. *Extract* N *values from the (parametric or empirical) unconditional distribution of the idiosyncratic disturbance,* $\tilde{\boldsymbol{u}}_0$*.*
5. *Construct the error terms* $\tilde{\boldsymbol{e}}_0 = \tilde{\boldsymbol{\alpha}} + \tilde{\boldsymbol{u}}_0$ *and order them from high to low.*
6. *Go down the ranking of the score (from high to low) and assign to each individual with* $y_0 = 0$ *the lowest unassigned value* $\tilde{e}$ *such that* $\tilde{e} < \boldsymbol{x}'_{j,0}\hat{\boldsymbol{\beta}}$*.*
7. *Go up the ranking of the score (from low to high) and assign to each individual with* $y_0 = 1$ *the highest unassigned value* $\tilde{e}$ *such that* $\tilde{e} \geq \boldsymbol{x}'_{j,0}\hat{\boldsymbol{\beta}}$*.*

8. *Go down the ranking of the score again (from high to low) and assign to each unmatched individual with* $y_0 = 0$ *the lowest unassigned value* $\tilde{e}$.
9. *Go up the ranking of the score again (from low to high) and assign to each unmatched individual with* $y_0 = 1$ *the highest unassigned value* $\tilde{e}$.

Steps 6-7 assign the random terms in such a way to maximize the number of concordant outcomes (observed and predicted): for those individuals with $y_0 = 0$, higher values of the score (which would by themselves increase the likelihood of a success) are matched with lower residuals; however, residuals that are not small enough to bring the value of the predicted latent variable below 0 are not "wasted" and are kept for observations with a lower value of the score. Analogously, for those individuals with $y_0 = 1$, lower values of the score (which would by themselves increase the likelihood of a failure) are matched with higher residuals; however, residuals that are not big enough to bring the value of the predicted latent variable above 0 are not "wasted" and are kept for observations with a higher value of the score. After this first round of assignments, there could be some observations and residuals still unmatched: these are the individuals for whom no matter the value of the residual left unassigned —given the extracted vector $\tilde{\boldsymbol{e}}_0$— the predicted outcome is different from the observed one. Therefore, the loop is repeated, without controlling for the threshold condition (steps 8-9).

C.1 Properties of the Conditional Rank method, binary case

In the binary response case, optimality of the Conditional Rank method in terms of minimization of MSFE($\boldsymbol{y}_0$) holds almost by construction.[23]

Proposition 5. *Given the model* (3)*, the solution obtained by applying algorithm 2 minimizes MSFE*$(\boldsymbol{y}_0) = \sum_j (y_{j,0} - \tilde{y}_{j,0})^2/N$.

[23]Note however that in the binary response case there might be more than one solution to the optimal assignment problem (for instance, the algorithm can in principle produce two different results depending on whether $y = 0$ or $y = 1$ observations are processed first; also, arbitrarily swapping matches created in the second round of assignment —concerning discordant individuals— leaves the forecasting error unaffected).

Proof. In order to improve on the MSFE, positive residuals are useful only to bring negative predicted scores in the positive camp, for individuals with $y_0 = 1$. Conversely, negative residuals are useful only to bring positive predicted scores in the negative camp, for individuals with $y_0 = 0$. There is no value in keeping a big residual for an individual with $y_0 = 0$, or a small residual for an individual with $y_0 = 1$. At the same time, there is no value in assigning a big residual to a small predicted score, given $y_0 = 1$, if the sum does not turn positive, nor assigning a small residual to a big predicted score, given $y_0 = 0$, if the sum does not turn negative.

The Conditional Rank method closely follows these principles. Denote as a 'good' match the case when a residual is assigned to an individual such that the predicted and the observed outcome for that individual are concordant. In the first round of the imputation process, the highest possible number of good matches are created for $y_0 = 0$ individuals. Conditional on this, the highest possible number of good matches are then created for $y_0 = 1$ individuals. We have therefore to show that it is not possible to create a good match for an individual who is left unassigned after the first round, without destroying at least another good match.

Consider a $y_0 = 0$ individual who is left unassigned after the first round: this means that there are no residuals small enough for him, without destroying some other good match involving a $y_0 = 0$ individual with a higher score. To re-establish such a good match we would need to destroy another good match involving a $y_0 = 0$ individual with an even higher score. At some point, no such good matches will be available: finding a good match for the first unassigned individual leads to no improvement in the forecasting error.[24]

Consider now a $y_0 = 1$ individual who is left unassigned after the first round: this means that there are no residuals big enough for him, without destroying some other good match. Such a good match involves either a $y_0 = 1$ or a $y_0 = 0$ individual. In the first case, an argument symmetric to the one above can be made, and there is no

[24]The only effect is to replace a concordant individual which is further away from the tipping point with a concordant individual which is closest to the threshold —a case when a forecasting error is arguably more acceptable.

improvement in the forecasting error. In the latter case, exactly the same argument above can be invoked, and again there is no improvement in the forecasting error. □

C.2 Example

Application of the Conditional Rank method to the example of tables 1 and 2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
id	$\boldsymbol{x}_0\boldsymbol{\beta}$	$\boldsymbol{e}_0$	$\boldsymbol{y}_0$	$\mathrm{r}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}})$	$\tilde{\boldsymbol{e}}_0$	$\mathrm{r}(\tilde{\boldsymbol{e}}_0)$	$\tilde{\boldsymbol{y}}_0$
1	-2.79	-0.51	0	10	-1.16	7	0
2	-0.50	-0.64	0	7	-1.81	8	0
3	-0.37	-0.47	0	4	-2.37	9	0
4	0.36	-0.90	0	1	-2.54	10	0
5	-1.38	1.77	1	9	-0.43	4	0
6	-0.77	2.62	1	8	-0.60	5	0
7	-0.49	2.24	1	6	0.58	1	1
8	-0.39	1.70	1	5	-1.08	6	0
9	-0.05	0.32	1	3	0.24	2	1
10	0.21	0.83	1	2	0.12	3	1

Note: It is assumed that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$

Table 3: Optimal assignment, binary response case.

The first part of the table contains the projections when $y = 0$: all failures are matched. Individual *4* has the highest value of the score, both in the $y = 0$ group and in the whole sample. Given that her observed outcome is 0, she is matched with the lowest residual, among those extracted, which turns out to be low enough to bring the value of the predicted score below 0. Accordingly, the second highest score in the $y = 0$ group (individual *3*) is matched with the second lowest residual, the third highest (individual *2*) score with the third lowest residual, and so on. All these residuals are small enough to keep their matched individuals in the $\tilde{y} = 0$ camp.

The second part of table refers to the case $y = 1$: the goal is now taking the highest number of predicted score above 0. Note that the lowest observed score in this group (individual *5*) is matched only with the fourth highest residual. This is because there are no residuals whatsoever that are big enough to compensate for the low value of the observables, for this individual. The big residuals are therefore saved and assigned to those individuals for which they make a difference: the highest residual (.58) is assigned

to individual *7*, who has a value of the score of -.49; the second highest residual (.24) is assigned to individual *9*, who has a value of the score of -.05; while the third highest residual (.12) is assigned to individual *10*, with a value of the score that is already above 0.

While the first round of imputation (steps *6.* and *7.* in the algorithm) selects the concordant units (for which the predicted outcome is equal to the observed outcome), the second round (steps *8.* and *9.*) deals with the discordant ones. Such discordant units (individuals *5*, *6* and *8*) are assigned a residual such that those with a lower score get again a higher residual: this is consistent with the fact that we observe a success for these individuals (and indeed, should they have an observed outcome $y = 0$, the reverse would be true).

Note that the above example is purportedly built in order to show the possibility to have discordant units. In practice however, this is a rare case, and the forecasting error $\text{MSFE}(\boldsymbol{y}_0) = \sum_j (y_{j,0} - \tilde{y}_{j,0})^2/N$ is generally very small. In 100 Montecarlo experiments with $x \sim N(0,1), \beta = 1, A \sim N(0,1), U \sim N(0,1)$ —supposing again that β is perfectly estimated— and a population of 100 individuals, the fraction of discordant units (unassigned residuals after the first round of imputation) is on average equal to 0.01%. Lowering the relevance of UH, for instance by having the explanatory variable extending over a wider range or centering its distribution further away from the tipping threshold for determining the outcome, has little effects: with $x \sim N(0,5)$ the fraction of discordant units is 0.03%; with $x \sim N(-.5,1)$ the fraction of discordant units is zero.

The Conditional Rank method is equivalent to the Hungarian method developed in Linear Programming, meaning that they find exactly the same solutions. However, it is faster: while the Hungarian method works, in the best operationalizations, in cubic time, the algorithm described here works in quadratic time. This is shown in figure 10, which plots the CPU usage of the Conditional Rank method for increasing values of the population size N, and $x \sim N(0,1), \beta = \hat{\beta} = 1, A \sim N(0,1), U \sim N(0,1)$.

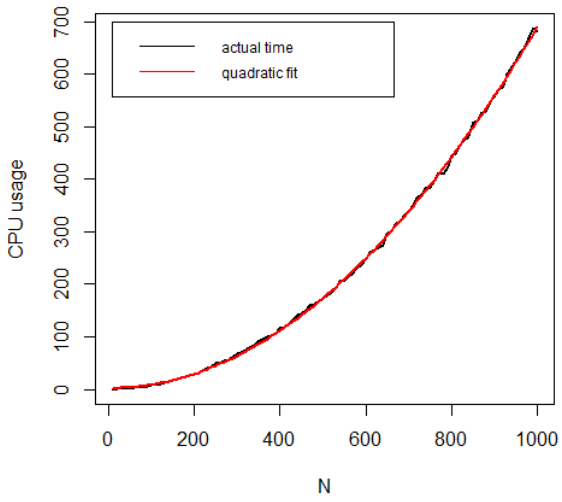


Figure 10: Speed of execution as a function of sample size, optimal assignment, binary response model. A quadratic fit is superimposed. Parameterization: $x \sim N(0,1), \beta = \hat{\beta} = 1, A \sim N(0,1), U \sim N(0,1)$.

References

Azevedo-Filho, A. and Shachter, R. D. (1994). Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In *Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference*, pages 28–36, San Mateo, CA. Morgan Kaufmann.

Bolstad, W. (2010). *Understanding Computational Bayesian Statistics*. John Wiley.

Burkard, R., Dell'Amico, M., and Martello, S. (2012). *Assignment Problems - Revised Reprint*. Society for Industrial and Applied Mathematics, Philadelphia.

Carpaneto, G., Martello, S., and Toth, P. (1988). Algorithms and codes for the assignment problem. *Annals of Operations Research*, 13:193–223.

Casella, G. and George, E. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Gu, Y., Fiebig, D. G., Cripps, E., and Kohn, R. (2009). Bayesian estimation of a random effects heteroscedastic probit model. *The Econometrics Journal*, 12:324–339.

Howell, D. C. (2008). The treatment of missing data. In Outhwaite, W. and Turner, S., editors, *The SAGE Handbook of Social Science Methodology*. SAGE publications.

Koopmans, T. and Beckmann, M. (1957). Assignment problems and the location of economic activities. *Econometrica*, 25(1):53–76.

Laplace, P.-S. (1774). Memoire sur la probabilite des causes par les evenements. english translation by s.m. stigler in 1986 as "memoir on the probability of the causes of events" in statistical science, 1(3), 359-378. *L'Academie Royale des Sciences*, 6:621–656.

Laplace, P.-S. (1814). *Essai Philosophique sur les Probabilites." English translation in Truscott, F.W. and Emory, F.L. (2007) from (1902) as "A Philosophical Essay on Probabilities". ISBN 1602063281, translated from the French 6th ed. (1840).*

Li, J. and O'Donoghue, C. (2013). A survey of dynamic microsimulation models: uses, model structure and methodology. *International Journal of Microsimulation*, 6(2):3–55.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley & Sons.

Moon, H. R., Perron, B., and Phillips, P. C. (2014). Incidental parameters and dynamic panle modeling. In *Handbook of Panel Data*.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

Panis, C. W. (2003). Microsimulations in the presence of heterogeneity. University of Michigan Retirement Research Center Working Paper 2003-048, University of Michigan.

Pickles, A. (2005). Missing data, problems and solutions. In Kempf-Leonard, K., editor, *Encyclopedia of Social Measurement*. Elsevier.

Richiardi, M. G. and Poggi, A. (2014). Imputing individual effects in dynamic microsimulation models. an application to household formation and labor market participation in italy. Technical report, Revised and resubmitted to the International Journal of Microsimulation.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–92.