Erasmus University Rotterdam

Master thesis Econometrics & Management Science

# Modelling hemoglobin levels of blood donors

Jesse Fokkinga (381822)

October 16, 2018

**Abstract**

To protect blood donors from anemia, donors with low hemoglobin levels are generally deferred from donating. In this thesis we consider various statistical techniques to gain insight in the trajectory of hemoglobin and to predict future values, which both can be useful to prevent deferrals. We examine the longitudinal association of hemoglobin (Hb) and zinc protoporphyrin (ZPP), a biomarker which is believed to be predictive for future Hb levels. We apply a multivariate autoregressive mixed-effects model, and find that our data suggest that there is not a time-dependent association, but rather a correlation of individual specific average values of Hb and ZPP. In the context of out-of-sample predictions, the usefulness of ZPP as a predictor for future Hb levels seems very limited. In order to successfully predict future Hb levels we examine a variety of methods that can be used for longitudinal forecasting. We propose a hierarchical specification of the mean-reverting Ornstein-Uhlenbeck process model, which matches well with the theoretical properties of the trajectory of Hb levels, and we provide a way to generate dynamic predictions with this model. Furthermore, we implement a Bayesian variable selection technique to incorporate an additional relatively high-dimensional set of blood levels, and we consider two decision tree ensemble methods to capture possible non-linearities in our data. We focus on out-of-sample forecasting and find that (i) the hierarchical Ornstein-Uhlenbeck model performs slightly better than traditional mixed-effects models, (ii) the tree based methods seem to be most successful in determining eligibility for donation and (iii) incorporating additional blood levels can improve predictions.

**Keywords:** Bayesian methods, multivariate autoregressive mixed-effects model, initial conditions problem, decision tree learning, Ornstein–Uhlenbeck process, spike and slab regression, blood donations, hemoglobin

Supervisor Erasmus MC: dr. Joost van Rosmalen
Supervisor Erasmus School of Economics: prof. dr. Richard Paap

# 1 Introduction

Blood transfusion is an essential part of modern healthcare, which helps to save a large number of lives every single day. Even though the field of medical sciences has made great progress in various directions over the last decades, an artificial alternative for real blood is not yet available. Therefore, the supply of blood is still fully dependent on blood donors. There are several types of blood donations, but the most common type is whole blood donation. Whole blood donation refers to a standard (usually 500 ml) donation, with other possible types of donation being plasma and platelet donation. A potential harm for whole blood donors is that donations could cause a loss of iron and blood cells, putting the donors at risk for anemia. In order to mitigate that risk, hemoglobin (Hb) concentrations are measured prior to every donation and donors with Hb levels below a certain threshold are deferred from donation.

Hemoglobin is a protein found in red blood cells that facilitates oxygen transport in the body of vertebrates. Hb is a biomarker that is very frequently used to assess the overall health condition of an individual, and decreased Hb levels provide an indication for anemia. As iron is the most important element of the Hb protein, a loss of iron can lead to a decrease of Hb levels. Insufficient consumption or excessive loss of iron can lead to iron deficiency and eventually anemia can develop. Anemia can have many different causes, including vitamin deficiencies, chronic diseases, but also excessive blood loss because of whole blood donations. Potential symptoms of anemia include fatigue, shortness of breath, impairment of cognitive functions and several other long term risks (Janz et al., 2013).

As Hb levels are impacted by donating blood, the Dutch blood bank Sanquin imposed a minimum time of 56 days between two successive donations. The idea behind imposing a minimum time between two donations is that the Hb levels of donors are impacted by a donation. On average a donation of 500 ml of blood causes male and female donors to lose 242 and 217 mg of iron respectively (Simon, 2002). This will cause Hb to decrease and reach its lowest value a few days after donation. The body will then start to reproduce Hb, whereupon the Hb level will gradually recover to its predonation value (Kiss et al., 2015 and Boulton, 2004). The required 56 days between two donations are assumed to be sufficient for Hb levels to fully recover to their initial values, but still a considerable proportion of blood donors is deferred from donation each year due to their Hb levels being too low. These deferrals can be costly, because a new donation needs to be scheduled. More importantly, it is demotivating for the donor and the probability of a donor not returning is relatively high.

In order to reduce deferrals, it can be useful to generate predictions of Hb levels and assess the predictive power of variables that can possibly explain (future) Hb levels. In this thesis we use statistical models to focus on two research objectives. Firstly, we examine the longitudinal association of Hb and a biomarker named zinc protoporphyrin (ZPP), which is a compound found in red blood cells and is believed to be predictive for Hb levels. Secondly, we evaluate strategies to obtain accurate Hb predictions. For both research objectives we start with the relatively basic univariate autoregressive mixed-effects model (Diggle et al. (2002)). In order to further examine the longitudinal association of ZPP and Hb we apply a multivariate autoregressive mixed-effects model. That is, we use a generalization of the usual vector autoregression model (Heij et al. (2004)) by embedding it in the linear mixed-effects framework (Verbeke (1997)). In order to predict future Hb levels as well as possible, we apply three types of techniques that can be used for longitudinal forecasting. We examine the performance of (i) two decision tree based ensemble methods, (ii) a newly proposed hierarchical

specification of the Ornstein-Uhlenbeck process model and (iii) a spike and slab regression model.

In the research of Baart et al. (2013) it is concluded that ZPP measurements have added value in the prediction of future Hb levels. In clinical practice, measurements of ZPP in red blood cells are already used as a screening test for lead poisoning and iron deficiency (Crowell et al., 2006 and Martin et al., 2004). The hypothesized association between ZPP and Hb is as follows: ZPP levels start to increase in the early stage of iron deficient erythropoiesis[1]. More specifically, ZPP is formed during heme synthesis[2] in case of iron deficiency. When iron levels are low, more zinc rather than iron is incorporated into protoporphyrin IX during the heme synthesis. This results in the formation of more ZPP and less heme, and as a result, ZPP accumulates in the blood cells. This implies ZPP measurements can detect iron deficiency in an early stage before Hb levels decrease (Baart et al., 2013). Therefore, biological theory suggest ZPP could be useful for predicting future Hb levels. In order to test this theory, we employ a multivariate autoregressive mixed-effects model, which is able to identify different types of association structures of the two biomarkers.

In the context of Hb level predictions, we argue that traditional methods that dominate current applied biostatistical research can potentially be improved. Firstly, we propose a novel hierarchical specification of an Ornstein-Uhlenbeck model, which theoretically matches well with the theoretical properties of Hb levels and is able to explicitly account for the fact that the measurements in our data are unequally spaced. Secondly, we examine the added value of two decision tree based ensemble methods. These methods can potentially further improve performance due to their increased flexibility and their ability to account for any nonlinearities in our data. Lastly, we implement a Bayesian variable selection technique named spike and slab regression in the context of a mixed-effects model specification in order to exploit a relatively high-dimensional set of additional blood levels, which could possibly be predictive for Hb levels.

Earlier research on statistical models for Hb levels is given by Nasserinejad et al. (2013), in which mixed-effects and transition (autoregressive) models were applied for the prediction of Hb levels. The results indicate that the transition model provides somewhat better predictions than the mixed-effects model, especially at a higher number of visits. Furthermore, in Nasserinejad et al. (2016) a latent class model was applied to (i) examine the required time between two subsequent donations, and (ii) predict future Hb levels. The authors find that the estimated recovery time is longer than the current minimum interval between donations, suggesting that an increase of this interval may be warranted. This notion is also further examined in this thesis.

The rest of this thesis is organized as follows: in Section 2 we present the data that we will use in our research. In Section 3 we describe various statistical topics that are relevant for analyzing longitudinal data and we provide the basis for the methodology used in this thesis. In Section 4 we introduce the multivariate autoregressive mixed-effects model, which we will use to examine the longitudinal relation between ZPP and Hb. In Section 5 we will elaborate on the aforementioned methods with which we aim to improve the predictions of future Hb levels. In Section 6 we discuss the empirical results of our analyses. In Section 7 we conclude with a review of our findings and provide proposals for future research.

[1] Erythropoiesis is the process which produces red blood cells.

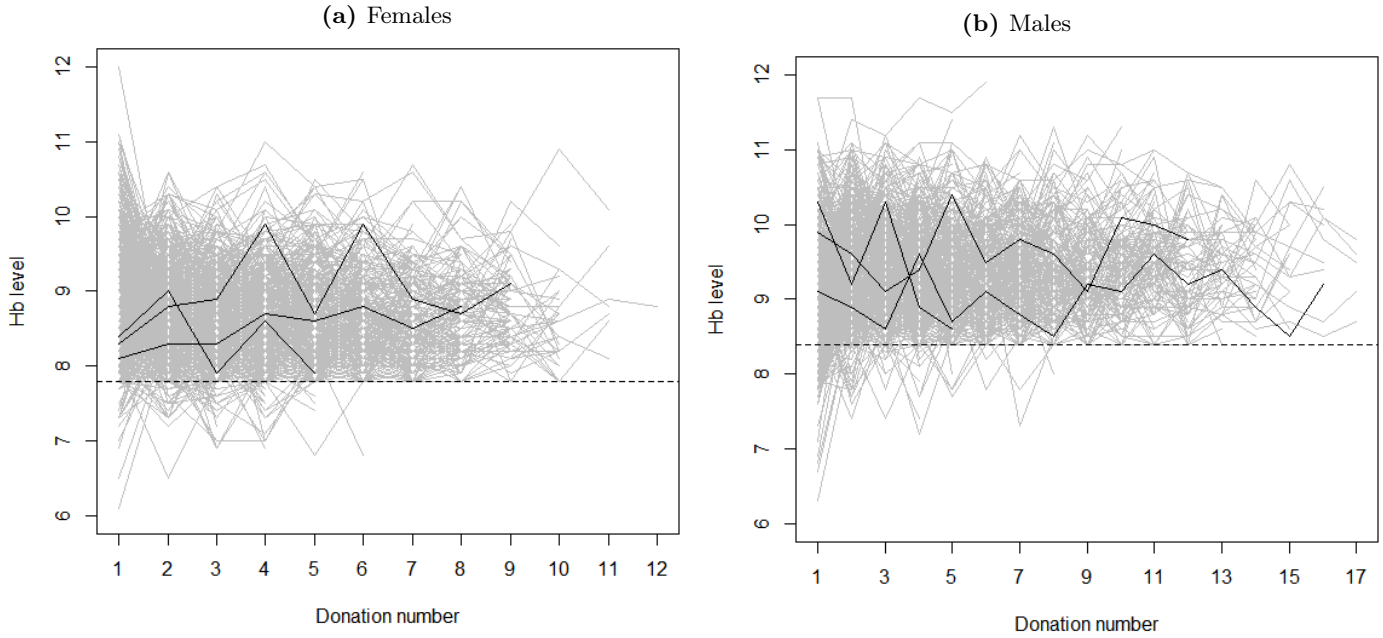
[2] Heme is an iron-containing compound which forms the non-protein part of hemoglobin and some other biological molecules.

## 2 Data

The data that we will use in this thesis has been gathered by Sanquin, which facilitates blood donations in the Netherlands. The data describes 14006 observations (donations) of 2215 different whole blood donors. The data has been collected during the period October 2009 - February 2014. Each visit includes a small screening prior to each donation. This screening involves taking a fingerstick capillary sample for measuring Hb level along with filling out a health appraisal form. Based on a too low Hb level or information on the appraisal form, an individual can be deferred from donation. The data used in this thesis has a longitudinal structure. That is, the data contains repeated observations of several donors where the number of observations of each specific donor differs and the observations are unequally spaced.

Because the data has not been gathered from an experimental study, there is no information about the Hb levels between subsequent donations; only measurements at the time of donation are available. Regarding the donation moments of individuals, there are two restrictions. Firstly, the minimum time interval between two successive donations should be at least 56 days. Secondly, the maximum number of yearly donations is 3 for females and 5 for males. Conditional on meeting these two requirements, individuals are free to choose their desired moment to donate blood.

**Figure 1:** Trajectory of Hb levels of male and female donors. Each line represents all visits of an unique donor. The profiles of 3 randomly selected donors that had at least three visits are highlighted. The dashed horizontal lines show the Hb cut-off values of eligibility for donation.




Several factors are known to be associated with Hb values, such as season, body mass index (BMI) and age (Yip et al., 1984 and Hoekstra et al., 2007), which implies these variables can be useful variables when modelling Hb levels. Because the trajectory of Hb is in general quite different for males and females and because the thresholds of eligibility for donation differ, we consider separate

models for males and females. When looking at the trajectory of Hb levels of the individuals in our data, as shown in Figure 1, we cannot identify any clear relation between the number of donations and Hb levels. The plots do confirm that women on average have lower Hb levels than males, and we can observe a clear between-individual variation of Hb levels. That is, some individuals have consistently higher Hb levels than others.

### 2.1 Explanatory variables

Our data contains information on several variables that can be used to explain and predict Hb levels. In our models we incorporate multiple explanatory variables that describe a specific donor. We use the age of a donor, Body Mass Index (BMI), estimated blood volume and an indicator whether a female is post menopausal[3]. For the blood volume we use the approximation by Nadler's formula (Nadler et al., 1962). That is, we approximate the blood volume of an individual based on its height and weight. Blood volume can potentially be an important predictor as an increased blood volume is typically associated with higher Hb levels. As additional explanatory variables we use season of donation, hour of the day, number of previous donations over the last two years, time since previous donations and previous ZPP value. The variable 'season of donation' describes in which season a specific donation took place, i.e. winter, autumn, summer or spring. The variable 'hour of the day' describes the hour at which the donation took place as a numeric value. For example, if a donation took place at 14:45 (2:45 PM), the value for this variable would be equal to 14.75. The full list of the explanatory variables that we use in our primary models is available in Table 1.

**Table 1:** List of explanatory variables which we include in our primary models

| **Variable** | **Type** |
|---|---|
| Season | Categorical |
| Post menopause | Dummy |
| Previous Hb measurement | Numeric |
| Previous ZPP measurement | Numeric |
| Estimated blood volume | Numeric |
| Body Mass Index | Numeric |
| Number of previous donations last 2 years | Numeric |
| Time since previous donation (in months) | Numeric |
| Time of the day (in hours) | Numeric |
| Age | Numeric |

For a subset of donations, we also have access to an additional set of blood levels. These blood levels contain information on i.a. platelets, red blood cell characteristics and the composition of white blood cells. The additional blood levels, along with measurements of ZPP, are derived from a lab analysis of the donated blood, and are thus not measured by means of a fingerstick capillary sample. We will specifically evaluate the benefit of exploiting these additional blood levels by incorporating these blood levels in a spike and slab regression model, which is described in Section 5.3. That is, we apply a specific regression model where we exploit the additional blood levels to predict Hb levels of individuals as our response variable, and we compare the predictive performance of this regression

[3]Post menopausal women have no menstrual flow. Note that this variable is not applicable for models describing male donors.

model to the performance of a regression model where we do not incorporate the additional blood levels. The full list of the blood levels is available in Appendix D.

**Table 2:** Descriptive statistics for the Sanquin data set

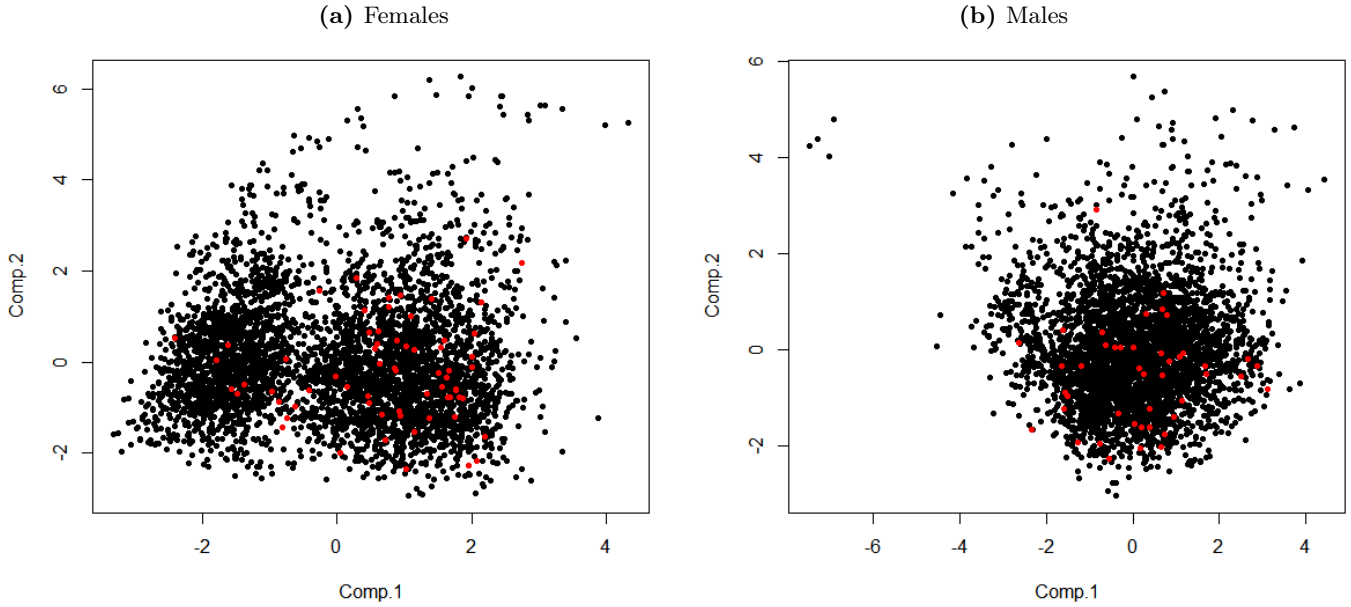
| | **Males** | **Females** |
|---|---|---|
| Age at 1st donations (years)* | 48 (34, 57) | 44 (31, 55) |
| Number of donations* | 6 (3, 8) | 3 (2, 5) |
| Time between donations (months)⊖ | 2.57 (1.46) | 4.45 (1.76) |
| Hb level⊖ | 9.32 (0.66) | 8.62 (0.62) |
| Log ZPP level⊖ | 4.05 (0.32) | 4.17 (0.33) |
| Number of donations for which full set of blood levels are observed | 1926 | 2191 |

* Median and inter-quantile range
⊖ Mean and standard deviation

In order to visually explore the structure of our data we apply Principal Component Analysis (PCA) to the set of explanatory variables of Table 1. PCA is a dimensionality reduction technique that aims to visualize data in a low-dimensional space by means of new variables that are a linear combination of the original variables. The results can be obtained by an eigenvalue decomposition of the covariance matrix of the data or a singular value decomposition of the data matrix. PCA generates components that are uncorrelated and explain as much of the variability of the original data as possible. A thorough explanation of PCA can be obtained from various sources such as Bro and Smilde (2014) and Jolliffe (2011).

The results of applying PCA to our data is shown in Figure 2. The plots show that it is very hard to separate specific groups of donors based on the first two principal components. For females we can identity a rough separation of two groups, which appears to be closely related to the menopausal status. In the right group, which approximately corresponds to visits of donors that still have menstrual periods, there is a larger number of deferrals. This is intuitive, as menstrual periods cause blood loss, which can lower Hb levels and therefore increase the risk of having low Hb levels. Despite this separation, we are not able to accurately identify clear groups of visits that have a larger risk of resulting in a deferral. It appears that for both males and females the linear combinations of our variables that explain the largest possible part of the variance in the data are unable to identify eligibility for donation.

**Figure 2:** Visualization of the data structure using the first two principal components (based on variables from Table 1). The red dots refer to visits that resulted in Hb measurements lower than the specified cut-off values of eligibility for donation.




# 3 Methodological framework

In this section, we elaborate on several statistical topics that are relevant for our research objectives of (i) examining the longitudinal association of Hb and ZPP and (ii) predicting future Hb levels. We briefly introduce the standard mixed-effects model and the mixed-effects transition model, which we use for the two research objectives. Afterwards, we explore the violated model assumptions that arise when combining a random intercept with an autoregressive term. We then describe our approach to estimate the statistical models in this thesis. Lastly, we provide a method to generate dynamic out-of-sample predictions with a random intercept model.

## 3.1 Models for longitudinal data

Hb levels are subject to both within-individual and between-individual variability. Two commonly used models to take into account within-individual correlation in a longitudinal setting are (i) mixed-effects models and (ii) transition (autoregressive) models. It can be useful to combine these two models in order to distinguish between unobserved heterogeneity and state dependence (Heckman, 1981) regarding the trajectory of Hb levels.

### 3.1.1 Linear mixed-effects model

The linear mixed-effects model is used extensively for analyzing longitudinal data (Verbeke, 1997). Apart from general parameters, the model also contains parameters that vary per individual. In our research we restrict ourselves to the inclusion of only an individual specific intercept, so that the

used model is defined as:

$$y_{it} = \alpha + b_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \tag{1}$$

$$\varepsilon_{it} \sim N(0, \sigma^2_\varepsilon), \qquad b_i \sim N(0, \sigma^2_b). \tag{2}$$

The vector[4] $\boldsymbol{x}_{it}$ refers to explanatory variables of individual $i$ at time $t$, with corresponding parameter vector $\boldsymbol{\beta}$. The random intercept $b_i$ can be interpreted as the deviation of the individual-specific mean of person $i$ from the population mean of Hb levels. This can be caused by several external factors that are not properly explained by the variables in the data, such as genetics, diet and exercise. The error term $\varepsilon_{it}$ is assumed to be independent of the random intercept. An important assumption for the consistency of the parameter estimates of the model is that the explanatory variables and the random intercept are independent (Hsiao, 2014).

In a linear mixed-effects framework, subsequent observations of a specific individual are marginally correlated, but are independent given the random effects:

$$p(\boldsymbol{y}_i|b_i) = \prod_{t=1}^{T_i} p(y_{it}|b_i) \tag{3}$$

Even though no explicit autocorrelation is modelled in a standard mixed-effects model, the model still allows for some correlation structure of subsequent observations of the same individual. This correlation is given by (Verbeke (1997)):

$$\rho(y_{it}, y_{it-1}) = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_\varepsilon}. \tag{4}$$

The linear mixed-effects model is a standard framework for longitudinal data analysis. For more information on the properties, extensions and practical usage of this type of model we refer to Verbeke (1997).

### 3.1.2 Mixed-effects transition model

Despite the fact that the mixed-effect model allows for some correlation structure of subsequent observations, it is unable to explicitly capture state dependence. State dependence occurs if the current outcome of a dynamic process depends on prior outcomes. In Heckman (1981), a strict distinction between true and spurious state dependence is made. In the context of biomarker values, true state dependence refers to a situation where a past biomarker level has a genuine effect on future values in the sense that the biomarker levels of a individual with a specific past biomarker level would behave differently in the future than those of an otherwise identical individual with a different past biomarker level.

True state dependence of two subsequent biomarker values is different from so called spurious state dependence. The latter refers to the case when individuals may differ in certain unmeasured variables that influence their biomarker value, while their current biomarker are not actually influenced by previous biomarker values. If the unmeasured variables are correlated over time, and are

[4]Throughout this thesis, we denote the transpose of a vector or matrix with an apostrophe. Thus, $\boldsymbol{x}'$ refers to the transpose of vector $\boldsymbol{x}$.

not properly controlled for, previous biomarker values may appear to be a determinant of future biomarker values only because they are a proxy for such temporally persistent unobservables.

In order to account for the true state dependence of Hb, we add an autoregressive term to our regression model. With the combination of a random intercept and an autoregressive term, we aim to distinguish between unobserved heterogeneity and true state dependence, which is captured by the autoregressive parameter. In particular, we consider the linear regression model where the response variable follows an AR(1) process:

$$y_{it} = \alpha + b_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \lambda y_{i,t-1} + \varepsilon_{it}, \tag{5}$$

where $b_i$ and $\varepsilon_{it}$ again follow normal distributions, as specified in Equation (2). After introducing the autoregressive term, the correlation between two successive observations is given by (Cameron and Trivedi (2005)):

$$\rho(y_{it}, y_{it-1}) = \lambda + \frac{1-\lambda}{1 + (1-\lambda)\sigma_\varepsilon^2/((1+\lambda)\sigma_b^2)}. \tag{6}$$

Equation (6) shows that if the transition effect is negligible, the correlation of two successive observations reduces to the intra-class correlation. Similarly, when the heterogeneity between individuals is very small ($\sigma_b^2 \approx 0$), the correlation is approximately equal to the transition effect.

In order to investigate the hypothesized association of ZPP and Hb, which suggests that ZPP levels could predict future Hb levels, the most simple approach would be to add the previous measurements of ZPP to the regression Equation (5) as extra a predictor, so that the estimated parameter for this predictor measures the effect of previous ZPP values on current Hb values. This can be achieved by fitting a regression model that corresponds to the equation:

$$y_{1it} = \alpha + b_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \lambda_1 y_{1it-1} + \lambda_2 y_{2it-1} + \varepsilon_{it}, \tag{7}$$

where $y_{1it}$ and $y_{2it}$ refer to measurements of respectively Hb and ZPP of individual $i$ at time $t$. Using this regression model, we can use the parameter estimates for $\lambda_2$ to examine whether ZPP can indeed explain future Hb values. As mentioned in the introduction, we also opt for using a multivariate approach to analyze the association. We will further elaborate on the multivariate model in Section 4.

### 3.2 Initial conditions problem

An important assumption of using a mixed-effects model is that the correlation between the random intercept and the predictors is zero, i.e., all the predictors are exogenous. In the mixed-effects transition model this assumption is violated, as one of the predictors is the lagged response variable, which is endogenous. This problem is known as the 'initial conditions problem' within the field of econometrics. When ignoring the initial conditions problem, the resulting parameter estimates are no longer consistent (Hsiao, 2014). In statistical literature, several solutions are proposed to correct for these violated model assumptions.

A well known approach to overcome the initial conditions problem is given by Heckman (1987). The idea of the approach is to jointly model an equation which links the random effects to the first observation of each individual:

$$y_{i0} = \boldsymbol{z}'_i\boldsymbol{\upsilon} + b_i\vartheta + \eta_i, \tag{8}$$

where $\boldsymbol{z}_i$ refers to exogenous variables that could be associated with the initial observation, and where $\boldsymbol{\upsilon}$ and $\vartheta$ are parameters corresponding to the exogenous variables and the random effects, respectively. Additionally, $\eta_i$ is the error term, which is assumed to be independent of the random effects. This linear specification, in terms of orthogonal error components, accounts for the possibility that the correlation between $y_{i0}$ and the random effects is non zero, which is expressed through the parameter $\vartheta$.

An alternative approach is given by Wooldridge (2005), who proposes a conditional maximum likelihood estimator that considers the distribution of the random effects conditional on the initial observation and exogenous variables. That is, we formulate the distribution of the random effects as:

$$b_i = \zeta_0 + \zeta_1 y_{i0} + \boldsymbol{z}_i' \boldsymbol{\upsilon} + a_i, \tag{9}$$

where $\boldsymbol{z}_i$ again refers to exogenous variables that could be associated with the initial observation, and $a_i$ is the normally distributed error term corresponding to the distribution of the random effects. We can now substitute the expression for $b_i$ of Equation (9) into our initial regression model equation, so that the term $a_i$ is now our new random intercept, which is assumed to be uncorrelated with the initial observation $y_{i0}$, thus satisfying our model assumptions. The solution of Wooldridge (2005) is easy to implement, as it only requires adding additional variables to the set of explanatory variables.

In Arulampalam and Stewart (2009) the two described solutions, along with a two-step estimator proposed by Orme (1996), were tested by means of an empirical simulation study. When comparing results obtained using these different correction methods, it was shown that all corrections yield similar results. This suggests that the choice which estimator to choose does not substantially impact estimate accuracy. Akay (2012) further investigates the Wooldridge solution and the Heckman solution under varying simulation setting. Based on the results of the simulations, the author claims Heckman's approach works better in shorter panels. As our dataset contains a large number of donors that donate infrequently during our observational period, we will implement the Heckman solution when estimating our models.

## 3.3 Bayesian inference

In order to estimate our models and generate forecasts, we opt for a Bayesian approach. Estimation of mixed-effect models is usually done in a frequentist framework by means of restricted maximum likelihood (Harville, 1977). However, there are several advantages for using a Bayesian approach for our analysis. First, maximizing the likelihood function of more complex mixed-effects models often brings computational difficulties, which can be overcome by relying on Bayesian inference. Second, Bayesian inference offers a natural approach to constructing mixed-effects model by means of hierarchical prior specifications, and it facilitates an intuitive way to make out-of-sample forecasts, which is further described in Section 3.4.1. Third, it allows to generate forecasts based on entire posterior distributions of the relevant model parameters, rather than relying solely on point estimates. Lastly, it provides exact inference in finite samples, without relying on asymptotic approximations.

### 3.3.1 Prior specification

We consider natural conjugate priors for all our parameters. For the variance of the random effects and the residuals, $\sigma_b^2$ and $\sigma_\varepsilon^2$, we impose inverse gamma priors. For each parameter in the full

parameter vector $\dot{\boldsymbol{\beta}} = (\alpha, \boldsymbol{\beta}', \lambda)'$, we consider a normal prior:

$$\begin{aligned} &\sigma_\varepsilon^2 \sim IG(\nu_\varepsilon, ss_\varepsilon), \quad \sigma_b^2 \sim IG(\nu_b, ss_b), \\ &\dot{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}_0, \boldsymbol{S}_0). \end{aligned} \tag{10}$$

Because we apply the Heckman solution to correct for the initial conditions problem, we jointly model the initial response of every individual. Therefore, we also have to specify prior distributions for the parameters corresponding to Equation (8). We again opt for conjugate prior specification for the parameters by imposing a normal prior for the regression parameters in $\dot{\boldsymbol{\upsilon}} = (\boldsymbol{\upsilon}', \vartheta)'$ and an inverse gamma prior for the residual variance $\sigma_\eta^2$:

$$\dot{\boldsymbol{\upsilon}} \sim N(\boldsymbol{\upsilon}_0, \boldsymbol{U}_0), \quad \sigma_\eta^2 \sim IG(\nu_\eta, ss_\eta). \tag{11}$$

As we have no prior information on our model parameters, we use diffuse priors for all our parameters. We implement these priors by setting $\nu_\varepsilon = \nu_\eta = \nu_b = 0.01$ and $ss_\varepsilon = ss_\eta = ss_b = 0.01$. For the prior of the regression parameters we set both $\boldsymbol{\beta}_0$ and $\boldsymbol{\upsilon}_0$ equal to a vector of zeros, and the prior covariance matrices $\boldsymbol{U}_0$ and $\boldsymbol{S}_0$ equal to diagonal matrices with $10^3$ on their diagonal, so that our prior means contain a lot of uncertainty and the predictors are assumed to be uncorrelated.

### 3.3.2 Posterior sampling

Due to the conjugacy of the priors, the conditional posterior distributions are of known form. Therefore, we can sample all model parameters by means of the Gibbs sampler. That is, we iteratively sample $\boldsymbol{b}$, $\dot{\boldsymbol{\beta}}$, $\sigma_\varepsilon^2$, $\sigma_b^2$, $\sigma_\eta^2$ and $\dot{\boldsymbol{\upsilon}}$ from their conditional distributions, which we can analytically derive from the full posterior distribution. Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm obtaining a sequence of observations which are sampled from a specific multivariate probability distribution. An extensive explanation of the Gibbs sampler can be obtained from several sources, such as Casella and George (1992).

## 3.4 Out-of-sample forecasting

In order to evaluate the out-of-sample prediction performance of our models we split the data in a train and test set. To split the data, we randomly select 90% of the individuals to be included in the train set, which we use to estimate our model parameters, and 10% of the individuals to be included in the test set, which we use to evaluate our model performance. We perform the random split between train and test, the estimation of our models and the prediction of observations in our test set 10 times. We then report the average performance over these 10 replications.

All of our fixed model parameters are estimated using data of observations in the train set, and the model predictions are evaluated using observations in the test set. In general, the predictive likelihood of observations in the test set is given by:

$$p(\boldsymbol{y}_{\text{test}}|\boldsymbol{y}_{\text{train}}, \boldsymbol{X}) = \int p(\boldsymbol{y}_{\text{test}}|\boldsymbol{\theta}, \boldsymbol{X}_{\text{test}}) p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{train}}, \boldsymbol{y}_{\text{train}}) d\boldsymbol{\theta}, \tag{12}$$

where the vector $\boldsymbol{\theta}$ contains all fixed model parameters. In order to obtain point predictions, which we use to evaluate our models, we consider the mean of the posterior predictive distribution $p(\boldsymbol{y}_{\text{test}}|\boldsymbol{y}_{\text{train}}, \boldsymbol{X})$. Note that because we use an autoregressive term in all of our models, we do not predict the first observation of an individual. As our models also contain an individual specific intercept, we need to separately estimate this parameter for individuals contained in the test set. The procedure of this estimation is explained in the section below.

### 3.4.1 Dynamic predictions

The forecasts of future Hb levels of a specific individual are based on fixed model parameters, the available covariates of an individual and on any previous donations an individual. After computing the posterior distributions for the fixed parameters in $\boldsymbol{\theta}$, we can dynamically 'update' the random intercepts of individuals in the test set whenever we gain new information. The posterior distribution for the random intercept is given by:

$$p(b_i^t|y_{it-1},...,y_{i1},\boldsymbol{X}_i,\boldsymbol{\theta}) \propto p(y_{it-1},...,y_{i1}|\boldsymbol{X}_i,b_i^t,\boldsymbol{\theta})p(\boldsymbol{b}), \tag{13}$$

where $b_i^t$ refers to the random intercept of individual $i$ at time $t$. The superscript $t$ is relevant, because $b_i^t$ is a dynamic parameter. That is, for the estimation of $b_i^t$ we can use all information available at time $t-1$, including all past observations of individual $i$. Therefore, for every new observation of individual $i$, we can update the random intercept. Because the model is linear in the random effects, the conditional posterior distribution is a normal distribution:

$$p(b_i^t|\boldsymbol{X}_i,\boldsymbol{\theta},y_{it-1},...,y_{i1}) = N(\bar{b_i},\sigma_{b_i}^2), \tag{14}$$

where $\bar{b_i}$ is the posterior mean and $\sigma_{b_i}^2$ is the posterior variance of the random intercept $b_i$. In order to find closed form expressions for $\bar{b_i}$ and $\sigma_{b_i}^2$, we first rearrange the terms from the initial observation, as given in Equation (8), and then divide all the terms by the standard deviation of the relevant error term:

$$(y_{i0} - \boldsymbol{z}_i'\boldsymbol{\upsilon})/\sigma_\eta = b_i\vartheta/\sigma_\eta + \eta_i/\sigma_\eta. \tag{15}$$

Similarly, subsequent observations of individual $i$ can be rewritten as:

$$(y_{it} - \boldsymbol{x}_{it}'\boldsymbol{\beta})/\sigma_\varepsilon = b_i 1/\sigma_\varepsilon + \varepsilon_{it}/\sigma_\varepsilon. \tag{16}$$

In order to express the sampling distribution of the random effects of individual $i$ at time $t+1$ we can stack the observations $0,...,t$ for every individual, resulting in the two auxiliary variables:

$$\boldsymbol{y}_i^* = \begin{pmatrix} (y_{i0} - \boldsymbol{z}_i'\boldsymbol{\upsilon})/\sigma_\eta \\ (y_{i1} - \boldsymbol{x}_{i1}'\boldsymbol{\beta})/\sigma_\varepsilon \\ \vdots \\ (y_{it} - \boldsymbol{x}_{it}'\boldsymbol{\beta})/\sigma_\varepsilon \end{pmatrix}, \quad \boldsymbol{x}_i^* = \begin{pmatrix} \vartheta/\sigma_\eta \\ 1/\sigma_\varepsilon \\ \vdots \\ 1/\sigma_\varepsilon \end{pmatrix}. \tag{17}$$

After stacking the observations of individual $i$, we can now use that $b_i$ is a parameter of an ordinary linear regression model with standard normally distributed error term. This conditional sampling distribution of the linear regression model is a well-known result, which is available in most introductory textbooks on Bayesian statistics such as Greenberg (2012). Due to the conjugacy of the prior specification $p(\boldsymbol{b})$, we now know that we can sample the random intercept of individual $b_i$ at time $t+1$ from a normal distribution with mean $(\boldsymbol{x_i^*}'\boldsymbol{x_i^*} + \sigma_b^{-2})^{-1}\boldsymbol{x_i^*}'\boldsymbol{y_i^*}$ and variance $(\boldsymbol{x_i^*}'\boldsymbol{x_i^*} + \sigma_b^{-2})^{-1}$. Note that in order to estimate the value of $b_i$ at time $t+1$ we can only use observations of individual $i$ up to time $t$.

The predictive distribution for observations in the test set can be decomposed as done in Equation (18). The predictions are dynamic because the predictive distribution of observations of an individual can be updated as soon as information from subsequent donations becomes available.

$$p(y_{it}|\boldsymbol{\theta},b_i^t,\boldsymbol{x}_{it}) = \int\int p(y_{it}|\boldsymbol{\theta},b_i^t,\boldsymbol{x}_{it})p(b_i^t|\boldsymbol{X}_i,\boldsymbol{\theta},y_{it-1},...,y_{i1})p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{train}},\boldsymbol{y}_{\text{train}})d\boldsymbol{\theta}db_i^t \tag{18}$$

### 3.4.2 Performance measures

In order to evaluate the performance of our models, we consider the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE).[5] One could also opt to compare models by means of measures like the deviance information criterion (Spiegelhalter et al., 2002) or the Bayesian information criterion (Schwarz et al., 1978), but in the context of predicting Hb levels the out-of-sample performance is more relevant than in-sample model fit.

In the Netherlands, the Hb thresholds of eligibility for donation are 7.8 and 8.4 mmol/l for women and men, respectively. In order to assess the performance of our models in terms of determining donation eligibility, we compute the receiver operating characteristic (ROC) curve for the predicted values of male and female donors. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold values. That is, the ROC curve evaluates a variety of values for which we classify a predicted donation to match the specified thresholds of 7.8 and 8.4. We then calculate the area under the curve (AUC) and use this measure to compare the models. The AUC measures discrimination, i.e., the ability of the models to correctly determine whether individuals are eligible for donation.

[5] We are aware that using the posterior mean of the predictive distribution to obtain point predictions makes the use of the RMSE the most logical approach; using the absolute error implies it would be natural to use the median of the predictive posterior distribution as a point prediction. Yet, in order to facilitate comparison to earlier research, we also use the MAE as an error measure.

# 4 Examining biomarker association with a multivariate autoregressive mixed-effects model

Rather than analyzing the hypothesized association of ZPP and Hb with an univariate regression model, as defined in Equation (7), a more sophisticated approach to modelling the association can be achieved by fitting a multivariate regression model. By using a multivariate model in which both ZPP and Hb are response variables, we can identify different types of association structures of the two biomarkers. As a result, we can hopefully reveal the true relation of ZPP and Hb that is present in our data. In particular, we propose to apply to following model:

$$\begin{pmatrix} y_{1,it} \\ y_{2,it} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} b_{1,i} \\ b_{2,i} \end{pmatrix} + \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix} \begin{pmatrix} y_{1,i,t-1} \\ y_{2,i,t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{x}'_{it}\boldsymbol{\beta}_1 \\ \boldsymbol{x}'_{it}\boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,it} \\ \varepsilon_{2,it} \end{pmatrix}, \tag{19}$$

$$(b_{1,i}, b_{2,i}) \sim N(\mathbf{0}, \boldsymbol{\Sigma_b}), \qquad (\varepsilon_{1,it}, \varepsilon_{2,it}) \sim N(\mathbf{0}, \boldsymbol{\Sigma_\varepsilon}), \tag{20}$$

$$\boldsymbol{\Sigma_b} = \begin{pmatrix} \sigma^2_{b1} & \sigma_{b12} \\ \sigma_{b21} & \sigma^2_{b2} \end{pmatrix}, \qquad \boldsymbol{\Sigma_\varepsilon} = \begin{pmatrix} \sigma^2_{\varepsilon 1} & \sigma_{\varepsilon 12} \\ \sigma_{\varepsilon 21} & \sigma^2_{\varepsilon 2} \end{pmatrix}. \tag{21}$$

This type of model is not used often in practice as of today in the field of biostatistics. In econometrics this type of model is known as a vector autoregression (VAR) model, while in psychology it is often referred to as a the cross-lagged panel model (CLPM). In biostatistics, the association between biomarkers or other outcomes is occasionally analyzed with multivariate mixed-effects models, which do involve random effects. However, to our knowledge the extension of also including autoregressive terms is still relatively uncovered.

Earlier literature on applications of multivariate mixed-effects models in the context of biostatistics include Oskrochi et al. (2016), where the model was used to investigate shoulder muscle functionality for patients that were treated for breast cancer. In Verbeke et al. (2014), a multivariate mixed-effect model was used to model loss of hearing ability. The authors treated measurements of hearing capabilities at different frequency levels as different response variables, and they specifically used the estimated covariance matrix of the random effects to make conclusions on the relation of hearing loss at different frequency levels. In econometrics, the combination of individual specific parameters and autoregressive terms in a multivariate regression model has been applied earlier by for example Fok et al. (2006), where the model was used to explain the differences in immediate and dynamic effects of promotional prices and regular prices on sales.

With the application of the multivariate autoregressive mixed-effects model we aim to empirically identify the 'true' association between ZPP and Hb that is present in our data. The multivariate autoregressive mixed-effects model is able to distinguish between three possible association structures of the two response variables:

- *Correlation of simultaneous outcomes.* This refers to the situations where correlation of contemporaneous measures of the two response variable exists. This can arise through the shared relation between the response variables and the observed predictors in the data or because of unobserved factors that impact both response variables simultaneously. The latter is expressed through the off-diagonal elements of $\boldsymbol{\Sigma_\varepsilon}$. Modelling the two outcomes separately in a univariate regression framework implicitly assumes these off-diagonal elements are zero.

- *Cross-lagged effects.* This association refers to the off-diagonal elements of the matrix $\boldsymbol{A}$. It could be possible that one response variable impacts the future path of the other response variable. That is, one response being high in a specific period indicates that the other response is going to be high or low in future periods and vice versa. This type of association relates to the theoretically assumed relation between ZPP and Hb, as described in the introduction.

- *Correlation of the random effects.* As described in Section 3.1.1, the random intercept can be interpreted as the deviation of the individual-specific mean of a response variable from the population mean of that response variable, which is due to unobserved factors such as genetics or diet. It could be possible that the deviation of the individual-specific mean from the population mean of one response variable for a specific individual indicates that in general the individual-specific mean of the other response variable of that individual also deviates from the population mean in a specific direction. If this is the case, the random intercepts of the two response variables are correlated. This association is expressed through the off-diagonal elements of $\boldsymbol{\Sigma_b}$. Again, when using an univariate regression model, these covariances are implicitly assumed to be zero.

When using an univariate regression model, we are able to only identify one of the three possible association structures. This can lead to invalid conclusions about the actual association structure of the two biomarkers and might not be able to determine the exact association that is present in our data. For inferential purposes it can be valuable to apply a multivariate autoregressive mixed-effects model, as this model is able to give a more extensive view on the longitudinal association of the two response variables.

## 4.1 Model estimation

As we use Bayesian inference for the estimation of the statistical models in our thesis, we have to specify priors for all model parameters associated with the multivariate autoregressive mixed-effects model. A complexity of the model is the combination of autoregressive terms and random intercepts, which again leads to the violated model assumptions, as earlier described in Section 3.2.

### 4.1.1 Initial conditions problem in a multivariate setting

In order to overcome inconsistency issues related to the initial conditions problem, we implement a multivariate generalization of the Heckman solution. That is, we jointly fit a separate multivariate equation which links the initial values of the two response variables to their respective random intercepts:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{z}'_{1i}\boldsymbol{\upsilon}_1 \\ \boldsymbol{z}'_{2i}\boldsymbol{\upsilon}_2 \end{pmatrix} + \begin{pmatrix} b_{1i}\vartheta_1 \\ b_{2i}\vartheta_2 \end{pmatrix} + \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \end{pmatrix}, \text{ with } \boldsymbol{\eta}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma_\eta}). \tag{22}$$

In current statistical literature, there is a very decent amount of information on the initial conditions problem available, along with relevant solutions and simulation studies empirically testing these solutions. Yet, to our knowledge there do not exist any simulation experiments that empirically test out any of these solutions in a multivariate context. The increased complexity of a multivariate model might further detriment the consistency of the parameter estimates or cause the established solutions to no longer function optimally. We will test out by means of a simulation study to what extent the Heckman solution for correcting for the initial conditions problem also works in a multivariate setting, and we will compare the results to the implementation of the Wooldridge solution.

As this simulation study might be tangential to the more applied nature of this thesis, the full design and the relevant results are available in Appendix A.

### 4.1.2 Prior specification

In our multivariate model, the variance parameters are no longer scalars. Hence, we use inverse Wishart priors for the covariance matrices of the residuals and the random effects. For the parameter vectors belonging to the two equations, we again consider normal priors. Using the notation that $\dot{\boldsymbol{\beta}}_i = (\alpha_i, A_{i,1}, A_{i,2}, \boldsymbol{\beta}_i')'$, we adopt the following prior specification:

$$\begin{aligned} &\boldsymbol{\Sigma_b} \sim IW(\boldsymbol{R_b}, \nu_b), \qquad \boldsymbol{\Sigma_\varepsilon} \sim IW(\boldsymbol{R_\varepsilon}, \nu_\varepsilon), \\ &\dot{\boldsymbol{\beta}}_1, \dot{\boldsymbol{\beta}}_2 \sim N(\boldsymbol{\beta}_0, \boldsymbol{S}_0) \end{aligned} \tag{23}$$

We also have to specify priors for the parameters relating to the regression Equation (22). Using that the parameter vectors for this equation are defined as $\dot{\boldsymbol{\upsilon}}_i = (\boldsymbol{\upsilon}_i', \vartheta_i)'$, we specify the following priors:

$$\begin{aligned} &\boldsymbol{\Sigma_\eta} \sim IW(\boldsymbol{R_\eta}, \nu_\eta), \\ &\dot{\boldsymbol{\upsilon}}_1, \dot{\boldsymbol{\upsilon}}_2 \sim N(\boldsymbol{\upsilon}_0, \boldsymbol{U}_0) \end{aligned} \tag{24}$$

As we have no prior knowledge, we consider a diffuse prior specification for all our model parameters. We set $\nu_\varepsilon = \nu_\eta = \nu_b = 3$ and we set the matrices $\boldsymbol{R}_\varepsilon, \boldsymbol{R}_\eta$ and $\boldsymbol{R}_b$ equal to $2 \times 2$ diagonal matrices with a $10^{-3}$ on their diagonal. These settings for the inverse Wishart distribution correspond to non-informative priors (Lesaffre and Lawson, 2012). In addition, we set $\boldsymbol{\beta}_0$ and $\boldsymbol{\upsilon}_0$ equal to vectors of zeros and we set the matrices $\boldsymbol{S}_0$ and $\boldsymbol{U}_0$ equal to diagonal matrices with $10^3$ on their diagonal.

# 5 Improving hemoglobin predictions

Current biostatistical methodology for predicting longitudinal outcomes is dominated by linear mixed-effects regression models. Even though in general these relatively straightforward models have proven to be successful for both inferential and predictive purposes, there are ways to potentially further improve the predictive accuracy. The traditional linear regression methods are unable to account for non-linear relations or interaction effects between variables. In order to overcome this issue, researchers can add non-linear transformations and interaction effects to their regression models (or splines, which are particularly popular in biostatistics). However, doing this forces the researcher to have some explicit knowledge or suspicion about the possible location of the non-linearities, which is often absent. In practice it is more desirable to have models that are able to identify these non-linearities themselves. For this purpose, we apply two decision tree based algorithms, which are described in Section 5.1.

Another downside of traditional methods is that typical linear regression models used for longitudinal modelling do not take the differences in time between subsequent measurements of the response variable into consideration. When adding an autoregressive term we do not explicitly account for the fact that the measurements are unequally spaced. Moreover, traditional biostatistical methods might not fit the characteristic properties of Hb levels of blood donors, which are theoretically assumed to be impacted by a donation, whereafter the levels gradually recover to their pre-donation value. In order to overcome these shortcomings, we propose an adapted version of the Ornstein-Uhlenbeck process model, which is described in Section 5.2.

As mentioned in Section 2, our dataset contains measurements of 21 variables describing blood levels. Initially, the idea of the researchers of Sanquin was to not use these variables, because of their relatively large quantity and the missing evidence that these variables are useful to specifically forecast Hb levels. Yet, from medical literature it is known that these variables can say something about the general health status of an individual, and therefore they might also be useful for predicting Hb levels. The relatively high dimension of this additional set of variables and the notion that these variables are likely to show a substantial degree of multicollinearity might make the use of traditional regression models not very appropriate. Instead, we rely on a Bayesian variable selection technique named spike and slab regression. In Section 5.3 we introduce the methodology behind this method. By using spike and slab regression, we can incorporate the additional variables in a more sophisticated manner than just adding them to a linear regression model.

## 5.1 Tree based methods

Decision tree learning uses a decision tree to exploit observations of several explanatory variables of a subject, which are represented the tree's branches, in order to make conclusions about the subject's target value, which is represented in the leaves of the tree. In our case we use regression trees to predict Hb levels as our target value. For a complete introduction on decision trees we refer to Tan et al. (2005).

The main motivation for applying decision trees to predict Hb levels is that decision trees are able to account for the possible non-linear relations and interaction effects of our explanatory variables. Tree based methods do not require any distributional assumptions, and they are intuitive to under-

stand for anyone with limited statistical knowledge. Moreover, algorithms to fit decision trees and several further extensions are widely available in software packages for the programming language R. Although decision tree based models do not involve an individual specific intercept which can account for unobserved heterogeneity, the models are able to accurately capture possible heterogeneity that is revealed through the observed covariates available in our data. As we only use the lagged dependent variables as a predictor in our models, we can no longer distinguish between true and spurious state dependence, as described in Section 3.1.2. However, as we do not use decision tree methods for any inferential purposes, this is not a major problem.

In general, single decision trees are known to be unstable and can be prone to overfitting (James et al., 2013). Ensemble methods aim to overcome this problem by aggregating the predictions of multiple trees, resulting in less volatile predictions. In order to improve prediction accuracy we use two well-known ensemble methods, which apply the general techniques of boosting and bagging to decision trees. Even though classification trees can also be used to predict deferrals as a binary outcome, we focus on using regression trees to predict Hb levels as a numeric outcome.

### 5.1.1 Random forest

Random forest is an ensemble method developed by Breiman (2001), in which the predictions of multiple individual decision trees are combined into a single prediction. The algorithm behind random forest applies the general technique of bootstrap aggregating to decision trees. Each single tree in the ensemble uses a different set of observations to generate predictions, as the observations for each decision tree are sampled with replacement from the full data set. In this thesis we use the Classification And Regression Tree (CART) algorithm, as proposed by Breiman et al. (1984), to fit single trees. The random forest algorithm can be summarized as follows:

1. Create $B$ subsamples by sampling with replacement from the full data
2. Fit a CART tree on each of the $B$ subsamples to obtain an ensemble of fitted decision trees
3. Use each individual tree in the ensemble to make a prediction of the response variable
4. Obtain the final prediction by taking the average prediction of the $B$ different decision trees

The idea behind the random forest algorithm is based on the paper of Breiman (1996), in which it is shown how bootstrap aggregating any predictor can improve prediction accuracy. The vital element is the instability of the prediction method. Bootstrap aggregating is able to significantly improve accuracy if perturbing the dataset that is used to train the predictor can cause significant changes in the constructed predictor. The somewhat unstable nature of decision tree algorithms makes that it is in general very useful to apply the technique of bootstrap aggregating to decision trees, leading to the motivation for the random forest algorithm.

An additional feature of the random forest algorithm is the parameter $h$, which is defined as the number of variables randomly sampled as candidates at each split. That is, when building the tree structure, the algorithm only evaluates $h$ of the total of $k$ explanatory variables as potential candidates to split on. In Breiman (2001) it is shown the accuracy of the random forest predictions is dependent on the strength of all $B$ individual tree and the correlation among the trees (less correlation leads to better predictions). The parameter $h$ provides a way to find the 'perfect balance'

between the two features. In general, a higher value for $h$ increases the strength of individual trees, but leads to a higher correlation, and vice versa. By tuning this parameter on the training data we can find the best value for $h$, which hopefully also leads to the best out-of-sample predictions.

### 5.1.2 Gradient tree boosting

The second decision tree ensemble method we consider is a relatively new adaptation of the idea of boosting in the context of decision trees, which has been proposed by Chen and Guestrin (2016). Again, this method uses an multitude of regression trees to make one single prediction. While in the case of random forests the trees are built independently of each other, in boosting each consecutive tree grown improves upon the previous tree. Trees in random forest are constructed by choosing splits that maximize impurity reduction, similar as in traditional decision tree algorithms. In gradient boosting, the splits are chosen such that a differentiable loss function is minimized. Moreover, the gradient boosting method contains various regularization parameters which can be tuned to decrease overfitting.

Traditional gradient boosting as proposed by Friedman (2001) is a regression method where new models are created that predict the residuals or errors of prior models and then are added together to make the final prediction. The algorithm that we use in this thesis optimizes an objective function $\mathcal{L}$ that takes decision trees as input. We aim to minimize the following objective:

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} l(y_{it}, \hat{y}_{it}) + \sum_{d=1}^{D} \Theta(f_d), \quad f_d \in \mathcal{F}, \tag{25}$$

where $\mathcal{F}$ is the function space of all tree functions. Each tree function is formally defined as $f_d(x) = w_{q(x)}, w \in \mathbb{R}^L, q : \mathbb{R}^p \to \{1, 2, .., L\}$. That is, a tree function $f_d(\boldsymbol{x}_{it})$ takes a set of variables $\boldsymbol{x}_{it}$ and maps this set to a leaf index $j = 1, 2, .., L$ corresponding to output value $w_j$ via the tree structure $q$. Equation (25) shows that the predicted values $\hat{y}_{it}$ are evaluated by means of a loss function $l$. This loss function can take on any form, such as the mean squared errors, which we will use as loss function. The second part, $\Theta(f_d)$, is the regularization term that controls the complexity of the model and thus prevents the model from overfitting. The regularization term is defined as $\Theta(f_d) = \gamma L + \frac{1}{2}\lambda \|\boldsymbol{w}\|^2$, where $L$ refers to the number of leaves in a tree and $\boldsymbol{w}$ refers to the weights on each of the leaves. Therefore, the parameter $\gamma$ penalizes the total number of leaves and the parameter $\lambda$ penalizes the leaf weights.

The loss function includes relatively complex tree functions as parameters and can not be optimized using traditional optimization methods. Therefore, our model is trained in an additive manner. That is, we add a new tree function each iteration of the algorithm. When we let $\hat{y}_{it}^{(d)}$ be the prediction of the Hb level of individual $i$ at time $t$ at iteration $d$, we will need to add a tree $f_d$ to minimize the following objective:

$$\mathcal{L}^{(d)}(q) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} l(y_{it}, \hat{y}_{it}^{(d-1)} + f_d(\boldsymbol{x}_{it})) + \Theta(f_d). \tag{26}$$

In order to find the optimal tree function at each iteration, the method proposed by Chen and Guestrin (2016) uses a second order Taylor polynomial approximation in order to quickly optimize

the objective:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n}\sum_{t=1}^{T_i}[l(y_{it}, \hat{y}_{it}^{(d-1)}) + g_{it}f_d(\boldsymbol{x}_{it}) + \frac{1}{2}h_{it}f_d^2(\mathbf{x}_i)] + \Theta(f_d), \tag{27}$$

where $g_{it} = \partial_{y^{d-1}} l(y_{it}, y_{it}^{(d-1)})$ and $h_{it} = \partial^2_{y^{d-1}} l(y_{it}, y_{it}^{(d-1)})$ refer to the first and second order gradient statistics on the loss function. The loss term $l(y_{it}, \hat{y}_{it}^{(d-1)})$ does not depend on the tree structure of iteration $d$ and can thus be considered a constant term. Therefore, we can rewrite a simplified approximated loss function as:

$$\tilde{\mathcal{L}}^{(d)} = \sum_{i=1}^{n}\sum_{t=1}^{T_i}[g_{it}f_d(\boldsymbol{x}_{it}) + \frac{1}{2}h_{it}f_d^2(\boldsymbol{x}_{it})] + \Theta(f_d). \tag{28}$$

Let $I_j = \{i|q(\boldsymbol{x}_{it} = j\}$ denote the set of observations that is mapped to leaf $j$. Using this notation, we can rewrite Equation (28) as:

$$= \sum_{l=1}^{L}\left[(\sum_{it\in I_j} g_{it})w_j + \frac{1}{2}(\sum_{it\in I_j} h_{it} + \lambda)w_j^2\right] + \gamma L. \tag{29}$$

For a fixed structure $q(\boldsymbol{x}_{it})$, we can now compute the optimal weight $w_j^*$ of leaf $j$ by:

$$w_j^* = -\frac{\sum_{it\in I_j} g_{it}}{\sum_{it\in I_j} h_{it} + \lambda}, \tag{30}$$

with associated optimal loss

$$\tilde{\mathcal{L}}^{(d)}(q) = -\frac{1}{2}\sum\nolimits_{j=1}^{L}\frac{(\sum_{it\in I_j} g_{it})^2}{\sum_{it\in I_j} h_{it} + \lambda} + \gamma L. \tag{31}$$

The optimal loss measures the quality of a tree structure in the gradient tree boosting algorithm, and is similar to impurity measures in traditional tree algorithms. In order to build the tree structure, the algorithm initializes with a single leaf and adds branches iteratively, a procedure which is identical to traditional decision tree algorithms. For a binary split, let $I_L$ contain all observations in the left set of nodes and $I_R$ contain all observations in the right set of nodes. Now, let $I = I_L \cup I_R$. The loss reduction after the split can be expressed as (Chen and Guestrin (2016)):

$$\frac{1}{2}\left[\frac{(\sum_{it\in I_L} g_{it})^2}{\sum_{it\in I_L} h_{it} + \lambda} + \frac{(\sum_{it\in I_R} g_{it})^2}{\sum_{it\in I_R} h_{it} + \lambda} - \frac{(\sum_{it\in I} g_{it})^2}{\sum_{it\in I} h_{it} + \lambda}\right] - \lambda. \tag{32}$$

This loss reduction is used to evaluate the loss reduction of potential splits, thus defining the way the tree structure is built. The parameter $\lambda$ is the minimum loss reduction required for a split. The algorithm of Chen and Guestrin (2016) contains an additional hyperparameter, the learning rate, by which all weights are shrunk after each iteration. This parameter helps to prevent the model from overfitting.

### 5.1.3 Hyperparameter tuning

Both tree ensemble methods that we use contain hyperparameters that need to be tuned before applying the methods to generate predictions. For the random forest algorithm, the only parameter which we will tune is $h$, which refers to the number of variables sampled as candidates at each split. In order to find the optimal value for $h$ we perform an exhaustive search combined with 10-fold cross validation on our test set. That is, we let $h$ range from 2 to $k-1$, where $k$ refers to the total number of variables. The training set is randomly partitioned into 10 equal sized subsamples, after which each single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used to train our model. We test out each single value of $h$, and eventually choose a value for $h$ for which the cross-validated prediction error is the lowest. Then, this value for $h$ is used to predict observations in the test set. We tune each of our model parameters based on the RMSE as the error measure.

The hyperparameter set-up of the gradient tree boosting algorithm is more complex when compared to random forest. The exact number of parameters to tune varies across different gradient boosting algorithms. We opt to fix certain parameters and only perform tuning on the parameters that get most attention in the literature. In order to find the optimal set of parameters without the risk of getting stuck at local optimal we perform a grid-search, again combined with 10-fold cross validation. In particular, we tune (i) the maximum tree depth, (ii) the learning rate, (iii) the fraction of observations that are randomly sampled to be used in each iteration and (iv) the fraction of variables that are randomly sampled to be used in each iteration. Note that this last parameter is very similar to the $h$ parameter of the random forest algorithm. As the grid-search is performed in a four-dimensional parameter space, the tuning of the hyperparameters related to the boosting algorithm is relatively time intensive. For larger datasets a more restricted way of choosing the optimal parameter values may be warranted.

For the purpose of determining whether donors will be eligible for donation, we can also choose to find the optimal value of our hyperparameters based on a different performance measure, such as the AUC. However, to facilitate comparison with our other prediction methods, in this research we solely tune the parameters based on the RMSE. As described in Section 3.4, we split the data in a train and a test set multiple times in order to evaluate the performance of our models. This means we also have to tune our parameters independently for each single train set.

## 5.2 A hierarchical approach to the Ornstein-Uhlenbeck model

The classic Ornstein-Uhlenbeck model describes a mean reverting stochastic process, which is based on a differential equation. The Ornstein-Uhlenbeck process is widely used for modelling several biological processes, and in mathematical finance, modelling the dynamics of interest rates and volatilities of asset prices. The Vasicek model, which is based on an Ornstein-Uhlenbeck process, describes the evolution of interest rates (Vasicek, 1977).

Treating Hb levels as an Ornstein-Uhlenbeck process is intuitive because of the characteristic properties of the trajectory of hemoglobin. When people donate blood, their Hb levels are believed to decline, after which they gradually recover to their initial values. These initial values can be described as the asymptotic means, to which the Hb levels gradually return as time passes by. Furthermore, the Ornstein-Uhlenbeck model explicitly takes into account the time that has passed

since the measurement of the previous Hb level, whereas the normal linear regression models just use an autoregressive term, neglecting the fact that the measurements are unequally spaced. The more time that has passed since a previous donation, the longer the donor has had to recover, and therefore the more likely it is that the Hb level will be close to its asymptotic mean.

### 5.2.1 Traditional model specification

The classic Ornstein-Uhlenbeck process satisfies the following stochastic differential equation:

$$dy_t = \kappa(\mu - y_t)\, dt + \tau\, dB_t, \tag{33}$$

with $\mu \in \mathbb{R}$, $\kappa, \tau \in \mathbb{R}_{>0}$ and $B(t)_{t\geq 0}$ being a standard Brownian motion. The parameter $\kappa$ can be interpreted as the decay-rate of the process, and describes how strongly the process reacts to perturbations. The asymptotic mean of the process is determined by the parameter $\mu$. The parameter $\tau$ describes the volatility of the process, and thus determines the size of the noise or variation.

Using Itô calculus, we can rewrite the Ornstein-Uhlenbeck process of random variable variable $y_t$ as:

$$y_t = e^{\kappa\delta_t} y_{t-1} + \mu(1 - e^{-\kappa\delta_t}) + \tau \int_0^t e^{\kappa(s-t)} dB_t, \tag{34}$$

where $\int_0^t e^{\kappa(s-t)} dB_t \sim N(0, \frac{1}{2\kappa}(1 - e^{-2\kappa\delta_t}))$ and $\delta_t$ refers to time between measurements $t$ and $t-1$. The only random component of the equation is the integral related to the Brownian motion. Substituting $\tau \int_0^t e^{\kappa(s-t)} dB_t = \varepsilon_t$, we can rewrite the process as a regression model:

$$y_t = e^{-\kappa\delta_t} y_{t-1} + \mu(1 - e^{-\kappa\delta_t}) + \varepsilon_t, \tag{35}$$

where $\varepsilon_t$ refers to the residuals of the regression equation, with $\varepsilon_t \sim N(0, \frac{\tau^2}{2\kappa}(1 - e^{-2\kappa\delta_t}))$.

### 5.2.2 Proposed adaptation to fit longitudinal data structure

We propose to adjust the traditional Ornstein-Uhlenbeck model to match the longitudinal structure of our data. We do this by estimating an individual specific asymptotic mean $\mu_i$, which we shrink towards a asymptotic population mean by using a hierarchical prior specification. Furthermore, we add a regression component to the expected value of the our response variable. In particular, our proposed hierarchical Ornstein-Uhlenbeck model is summarized as:

$$\begin{aligned} y_{it} &= e^{-\kappa\delta_{it}} y_{it-1} + \mu_i(1 - e^{-\kappa\delta_{it}}) + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \\ \mu_i &\sim N(\mu, \sigma_\mu^2), \qquad\qquad \varepsilon_{it} \sim N(0, \frac{\tau^2}{2\kappa}(1 - e^{-2\kappa\delta_{it}})), \end{aligned} \tag{36}$$

where $\boldsymbol{x}_{it}$ refers to the explanatory variables of individual $i$ at time $t$ and $y_{it}$ is the Hb level of individual $i$ at time $t$.

The downside of estimating an individual specific asymptotic mean is that we need to estimate a relatively large number of parameters, which is at the expense of the parsimony of the model. In order to mitigate this issue, we adopt a prior specification where we shrink the parameters towards a population parameter. The hierarchical prior allows the information in the data regarding the parameters describing the asymptotic mean of each individual to be shared. In general, hierarchical

priors are useful in more complex models where the parameter space is of such a dimension that the data is insufficient to properly identify all parameters. This is very applicable to our data, which contains only a limited number of donations for each individual.

### 5.2.3 Prior specification

As we are using a Bayesian approach to estimate our models and generate predictions, we have to specify priors for our model parameters. We adopt the following prior specification:

$$\begin{aligned} &\sigma_\mu^2 \sim IG(\nu_\mu, ss_\mu), \\ &\mu \sim N(\mu_0, \varphi_\mu^2), && \boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{S}_0), \\ &\tau \sim N(\tau_0, \varphi_\tau^2)\mathbb{1}_{[\tau>0]}, && \kappa \sim N(\kappa_0, \varphi_\kappa^2)\mathbb{1}_{[\kappa>0]}, \end{aligned} \tag{37}$$

where $\mathbb{1}_{[.]}$ denotes the indicator function, which equals one if the parameter in brackets is in the specified range and is zero otherwise. Note that we use truncated normal priors for $\kappa$ and $\tau$ because these parameters are restricted to be positive. As we have no prior knowledge, we opt for using diffuse priors for all our parameters. We do so by setting $\varphi_\sigma^2 = \varphi_\mu^2 = \varphi_\kappa^2 = 100$, $\nu_\mu = ss_\mu = 0.01$, $\mu_0 = \kappa_0 = \tau_0 = 0$, $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{S}_0$ equal to a diagonal matrix with $10^3$ on its diagonal.

### 5.2.4 Dynamic predictions

As we want to use the model to make out-of-sample predictions of future Hb levels, we need to dynamically update the individual specific asymptotic mean $\mu_i$ whenever we gain information on the trajectory of the Hb levels of individual $i$. We do this in a similar way as described in Section 3.4. Therefore, we need to derive the sampling distribution of the parameters that describe the individual specific mean. Rearranging the terms of Equation (36) and dividing every element by the standard deviation of $\varepsilon_{it}$, we obtain:

$$(y_{it} - e^{-\kappa\delta_{it}} y_{it-1} - \boldsymbol{x}'_{it}\boldsymbol{\beta})\tfrac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{it}})^{-1/2} = \mu_i(1 - e^{-\kappa\delta_{it}})\tfrac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{it}})^{-1/2} + \eta_{it}^*, \tag{38}$$

with $\eta_{it}^* = \varepsilon_{it}\frac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{it}})^{-1/2}$ being a standard normally distributed random variable. We can now define the auxiliary variables $\boldsymbol{y}_i^*$ and $\boldsymbol{x}_i^*$:

$$\boldsymbol{y}_i^* = \begin{pmatrix} (y_{i1} - e^{-\kappa\delta_{i1}} y_{i0} - \boldsymbol{x}'_{i1}\boldsymbol{\beta})\frac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{i1}})^{-1/2} \\ (y_{i2} - e^{-\kappa\delta_{i2}} y_{i1} - \boldsymbol{x}'_{i2}\boldsymbol{\beta})\frac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{i2}})^{-1/2} \\ \vdots \\ (y_{it} - e^{-\kappa\delta_{iT}} y_{it-1} - \boldsymbol{x}'_{it}\boldsymbol{\beta})\frac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{it}})^{-1/2} \end{pmatrix}, \tag{39}$$

$$\boldsymbol{x}_i^* = \begin{pmatrix} (1 - e^{-\kappa\delta_{i1}})\frac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{i1}})^{-1/2} \\ (1 - e^{-\kappa\delta_{i2}})\frac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{i2}})^{-1/2} \\ \vdots \\ (1 - e^{-\kappa\delta_{it}})\frac{\sqrt{2\kappa}}{\tau}(1 - e^{-2\kappa\delta_{it}})^{-1/2} \end{pmatrix} \tag{40}$$

Using these two variables and the conjugate prior specification of $\mu_i$, we use the standard results from an ordinary linear regression model to obtain the conditional sampling distribution of the individual specific parameter describing the asymptotic mean. That is, we can treat $\boldsymbol{y}_i^*$ as the dependent variable and $\boldsymbol{x}_i^*$ as the predictor of an ordinary linear regression model with standard

normally distributed error term in order to derive the sampling distribution of regression parameter $\mu_i$. Similar as described in Section 3.4, we know that the distribution of $\mu_i$ at time $t+1$ is a normal with mean $(\boldsymbol{x}_i^{*\prime}\boldsymbol{x}_i^* + \sigma_\mu^{-2})^{-1}(\boldsymbol{x}_i^{*\prime}\boldsymbol{y}_i^* + \mu\sigma_\mu^{-2})$ and variance $(\boldsymbol{x}_i^{*\prime}\boldsymbol{x}_i^* + \sigma_\mu^{-2})^{-1}$. Again, in order to estimate the value of $\mu_i$ at time $t+1$ we can only use observations up to time $t$.

### 5.3 Exploiting additional blood levels with spike and slab regression

In order to evaluate the benefits of the set of blood levels, which is described in Section 2, we apply spike and slab regression. Spike and slab regression is a Bayesian method that allows the exclusion of regressors whose coefficients are likely to be zero by means of hierarchical prior specification on the regression coefficients. By using spike and slab regression we are able to only select the variables that are relevant for predicting future Hb levels. Spike and slab regression allows us to (i) account for the uncertainty about whether each of the individual blood levels has added value when it comes to predicting hemoglobin, (ii) overcome the presence of multicollinearity amongst the variables and (iii) prevent the model from overfitting. Note that we also include our primary variables from Table 1 in our spike and slab regression model. We will examine the benefits of the blood levels by comparing the predictive performance of a spike and slab mixed-effects regression model in which we do incorporate the blood levels to the performance of a mixed-effects regression model without the blood levels. To avoid any look-ahead bias, we can only use the measured values of each of the blood levels at time $t-1$ to predict the Hb level of an individual at time $t$.

#### 5.3.1 Prior specification and posterior sampling

Spike and slab regression, as popularized by George and McCulloch (1997) and Ishwaran et al. (2005), uses a conjugate prior specification that imposes a normal mixture prior on the regression parameters $\boldsymbol{\beta}$ by introducing a binary parameter vector $\boldsymbol{\gamma}$ that defines which explanatory variables are included in the regression model. That is, the vector $\boldsymbol{\gamma}$ consists of zeros and ones, where $\gamma_i = 1$ indicates variable $i$ is included and $\gamma_i = 0$ implies exclusion. When one wants to apply variable selection on $k$ explanatory variables, any particular vector $\boldsymbol{\gamma}$ represents an unique regressor combination out of a total of $2^k$ possible combinations.

The prior specification for $\gamma$ is the product of $k$ independent Bernoulli variables, for which the prior inclusion probability $\varrho_i$ is allowed to vary per variable. The term 'spike and slab' originates from Mitchell and Beauchamp (1988) and refers to the prior specification of the model, which assumes a two-point mixture distribution made up of a uniform flat distribution (the slab) and a degenerate distribution at zero (the spike). The marginal distribution $p(\boldsymbol{\gamma})$ defines the spike, as it places probability mass at zero. The prior inclusion probability is usually determined on the basis of the expected model size $m$, with $\varrho = m/k$.

Conditional on $\boldsymbol{\gamma}$, the posterior distribution of $\boldsymbol{\beta}$ follows from the well-known posterior of an ordinary linear regression model with conjugate priors, with the only difference being the exclusion of variables whose value for $\gamma$ equals 0. The sampling procedure of the binary parameter vector $\boldsymbol{\gamma}$ is slightly more tricky. In George and McCulloch (1997) it is shown that the spike and slab algorithm improves by sampling from the marginal posterior of $\boldsymbol{\gamma}$. Although a closed-form expression for the marginal posterior of $\boldsymbol{\gamma}$ is not available, samples of $\boldsymbol{\gamma}$ can be constructed by means of an embedded Gibbs sampling routine that sequentially samples from the conditional Bernoulli distribution of $\gamma_i$ given $\boldsymbol{\gamma}_{-i}$ in a random order, where $\boldsymbol{\gamma}_{-i}$ refers to the set including all values of $\boldsymbol{\gamma}$ except for $\gamma_i$.

In essence, the algorithm evaluates for each variable $i$ the residual sum of squares and information matrix corresponding to $\gamma_i = 1$ and $\gamma_i = 0$. When a certain predictor has a large explanatory power, it will substantially reduce the residual sum of squares and it will receive a high probability of being included in the model. For a thorough explanation of the exact algorithm behind spike and slab regression with conjugate priors we refer to George and McCulloch (1997).

Apart from the Bernoulli prior for $\boldsymbol{\gamma}$, the conjugate prior set-up that we use is identical to the set-up described in Section 3.3.1. The only difference is that the prior for $\boldsymbol{\beta}$ depends on $\boldsymbol{\gamma}$, because the coefficients corresponding to a value of $\gamma = 0$ are defined to be equal to zero. Because we apply this method in the context of the mixed-effects transition model, we also implement the Heckman solution to the correct for the initials conditions problem. As we have no prior knowledge, we use diffuse prior settings, which are again identical to the settings described in Section 3.3.1. We set the prior inclusion probability equal to $\varrho = 0.25$.

### 5.3.2 Bayesian model averaging

In order to make out-of-sample predictions using a spike and slab regression model we use the same procedure as described in Section 3.4. The predictive distribution for observations in the test set is again given by Equation (18). In the context of spike and slab regression, integrating over $\boldsymbol{\theta}$ is a form of Bayesian model averaging, due to the parameters in $\boldsymbol{\gamma}$. Each draw of $\boldsymbol{\gamma}$ can essentially correspond to a different model, as in each iteration of the Gibbs sampler different predictors can be included in the model.

The idea behind Bayesian model averaging is that all predictors that are truly related to the dependent variable guide the posterior mean of the predictive distribution to its 'true value'. Contrarily, all predictors that are unrelated to the dependent variable, but do get included in the model at certain iterations of the Markov chain, individually push the mean away from the true value. Hopefully, these spurious inclusions are independent, so that they eventually average out to zero. In practice the Gibbs sampler might at some point get stuck in local modes of the parameter space. As a result, spurious relations between the predictor variables and the dependent variable may dominate the posterior. However, the risk of this problem is a lot higher in very high-dimensional regression problems. As in our data the number of explanatory variables is still relatively moderate, we expect that the algorithm is well able to eventually find the most ideal model specification.

# 6 Empirical results

In this section we present the results of applying the described methods to the data of Sanquin. We make a separation between the inferential results, in which we emphasize the longitudinal association of ZPP and Hb, and the results of the predictions of Hb levels. In order to improve model fit we normalize the skewed distribution of our data on ZPP by considering a log transformation. Note that for the inferential results we have used the full data to estimate the reported posteriors, while for our predictions we solely report the out-of-sample forecasting results.

In order to check for the convergence of our Markov chains, we run three different chains and examine standard Bayesian diagnostics, such as trace plots and Brooks-Gelman-Rubin statistics (Brooks and Gelman, 1998). We dismiss the first 5000 iterations of each chain, which proved to be sufficient to solely sample from converged chains. Only for the hierarchical Ornstein-Uhlenbeck model we had to chose a larger burn-in, which we set to 25000 iterations. In addition to the burn-in iterations, we used an additional 7000 draws to build our posterior distributions.

## 6.1 Inferential results

In order to analyze the relation of Hb and our explanatory variables, we consider the autoregressive mixed-effects model of Equation (7), with explanatory variables as described in Section 2. Additionally, we consider the multivariate mixed-effects model of Equation (19) using the same explanatory variables, but with both ZPP and Hb as outcome variables.

### 6.1.1 Univariate regression

**Table 3:** Posterior results of the regression analysis of Equation (7) with Hb as dependent variable

| | **Females** | | | **Males** | | |
|---|---|---|---|---|---|---|
| | mean | 95% CI | | mean | 95% CI | |
| (Intercept) | 7.855* | 7.426 | 8.297 | 8.119* | 7.570 | 8.670 |
| Age | 0.001 | −0.002 | 0.004 | −0.007* | −0.009 | −0.004 |
| Spring | 0.003 | −0.036 | 0.043 | −0.021 | −0.066 | 0.023 |
| Summer | −0.016 | −0.054 | 0.022 | 0.003 | −0.042 | 0.048 |
| Autumn | −0.009 | −0.049 | 0.031 | −0.026 | −0.072 | 0.020 |
| Time of the day (hours) | 0.000 | −0.005 | 0.004 | −0.006* | −0.011 | −0.001 |
| BMI | 0.005 | −0.004 | 0.013 | 0.017* | 0.005 | 0.029 |
| Bloodvolume | 0.053 | −0.018 | 0.123 | 0.041 | −0.028 | 0.111 |
| Number of previous donations last 2 years | 0.013 | −0.001 | 0.027 | 0.016* | 0.006 | 0.026 |
| Post menopause | 0.101* | 0.023 | 0.179 | - | - | - |
| Time since previous donation (months) | 0.001 | −0.003 | 0.006 | 0.007* | 0.001 | 0.013 |
| Previous Hb value | 0.104* | 0.071 | 0.136 | 0.137* | 0.101 | 0.173 |
| Previous ZPP value | −0.163* | −0.227 | −0.098 | −0.104* | −0.180 | −0.024 |

The asterisk (*) denotes that zero is not included in the 95% highest posterior density interval.

The posterior results in Table 3 suggest that previous ZPP values indeed have a significant[6] effect on measured Hb levels. Thus, according to the model it could be useful to exploit previous ZPP levels when generating predictions of Hb. This conclusion is in line with earlier findings by Baart et al. (2013). Another interesting finding is that the time since previous donation does not have appear to have a significant effect on Hb levels for females. Note that for each observation in our data, at least 56 days have passed by since the previous donation and there is maximum number of donations each year. Therefore, the lack of posterior support for this parameter suggests that extra time between two donations does not further impact the Hb levels. This implies that for females the current restrictions regarding donation moments are sufficient for Hb levels to fully recover.

The explanatory power of the seasonal indicators and the time of the day seems very limited. We also tried using trigonometric functions of the time of the day and the day of the year[7] in order to capture the impact of seasonality and timing of the donations, but this did also not result in any posterior support for a significant influence of these variables. The number of previous donations does not seem to significantly impact the Hb levels for females, while for males we even find posterior evidence for a positive effect, which is an unexpected result from a clinical perspective. The posterior distribution of the autoregressive parameter suggests that there is some degree of state dependence present in the data. Overall, the data suggests that the explanatory value of the majority of the predictors is low, resulting in lack of significant posterior evidence for many of our parameters.

### 6.1.2 Multivariate regression

The full results of fitting the multivariate regression model, which are shown in Appendix C, indicate that (i) the outcomes of ZPP seem to be more affected our explanatory variables than the outcomes of Hb, and (ii) the overall conclusion regarding the impact of the explanatory variables on Hb remains very similar. However, there is one eye-catching difference compared to the results from the univariate model: the effect of previous ZPP levels on Hb is no longer significant, and even switched sign for males. This result, which is shown in Table 4, can be explained by the fact that the multivariate model is able to account for (i) correlation of the random effects and (ii) correlation of the residuals, whereas our initial univariate model is not.

A closer look to the estimated covariance matrices of the random effects and the residuals suggests that it is mostly the correlation of the random effects that accounts for the disappearance of the cross-lagged effect. The posterior results of the covariance matrix of the random effects, as shown in Table 5, show an estimated correlation of the random effects of around -0.4. This finding, in combination with the absence of significant cross-lagged effects, suggests that the association of ZPP and Hb is mostly reflected through the correlation of the random effects.

In order to verify that the correlation of the random effects is responsible for the absence of the cross-lagged effect, we fit an additional multivariate model according to Equation (19), where we now

[6] With "significant effect" we mean that zero is not included in the 95% highest posterior density interval of the related parameter.

[7] For the seasonal component, we created two new variables by means of the functions $\sin(\frac{2\pi t}{T})$ and $\cos(\frac{2\pi t}{T})$, where $t$ denotes the day on which a donation took place and $T$ is the total number of days (365.25). Similarly, for the time of the day components we also created two variables by means of the functions $\sin(\frac{2\pi t}{T})$ and $\cos(\frac{2\pi t}{T})$, where now $t$ denotes the hour (as a numeric value) on which the donation took place and $T$ is the total number of hours (24). The idea for this approach originates from Stolwijk et al. (1999).

**Table 4:** Selection of posterior results of multivariate regression of Equation (19)

| Dependent variable | Predictor variables | **Females** | | | **Males** | | |
|---|---|---|---|---|---|---|---|
| | | mean | 95% CI | | mean | 95% CI | |
| Hb | Previous Hb | 0.094* | 0.063 | 0.125 | 0.130* | 0.095 | 0.165 |
| | Previous ZPP | −0.049 | −0.040 | 0.137 | 0.105 | −0.016 | 0.204 |
| ZPP | Previous Hb | 0.002 | −0.009 | 0.013 | 0.013 | −0.002 | 0.024 |
| | Previous ZPP | 0.213* | 0.175 | 0.252 | 0.241* | 0.201 | 0.281 |

The asterisk (*) denotes that zero is not included in the 95% highest posterior density interval.

**Table 5:** Posterior results of covariance matrix and correlation of the random effects belonging to the multivariate regression analysis of Equation (19)

| | **Females** | | | **Males** | | |
|---|---|---|---|---|---|---|
| | mean | 95% CI | | mean | 95% CI | |
| $\sigma^2_{b1}$ | 0.120 | 0.101 | 0.139 | 0.138 | 0.113 | 0.166 |
| $\sigma^2_{b2}$ | 0.046 | 0.040 | 0.053 | 0.038 | 0.031 | 0.044 |
| $\sigma_{b21}$ | −0.029 | −0.038 | −0.020 | −0.030 | −0.040 | −0.020 |
| $\rho(b_1, b_2)$ | −0.380 | −0.475 | −0.279 | −0.407 | −0.512 | −0.296 |

restrict $\boldsymbol{\Sigma_b}$ to be diagonal. This is equivalent to assuming that the random intercepts corresponding to Hb and ZPP come from independent normal distributions:

$$b_{i1} \sim N(0, \sigma^2_{\boldsymbol{b}1}), \qquad b_{i2} \sim N(0, \sigma^2_{\boldsymbol{b}2}). \tag{41}$$

The results of fitting these models with the restriction that no correlation of the random effects can exists, as partially shown in Table 6, confirm the notion that the absence of significant cross-lagged effects in our 'normal' multivariate model is caused by the fact that this model allows for presence of correlation of the random effects. In Table 6 we find results that are very similar to the results of the univariate models, which indicates that when we do not allow the random effects to be correlated, the off-diagonal element of parameter matrix $\boldsymbol{A}$ from Equation (19) again 'absorb' the association of the two response variables.

**Table 6:** Selection of posterior results of multivariate regression of Equation (19) with $\boldsymbol{\Sigma_b}$ restricted to be diagonal

| Dependent variable | Predictor variables | **Females** | | | **Males** | | |
|---|---|---|---|---|---|---|---|
| | | mean | 95% CI | | mean | 95% CI | |
| Hb | Previous Hb | 0.103* | 0.071 | 0.137 | 0.136* | 0.101 | 0.172 |
| | Previous ZPP | −0.164* | −0.228 | −0.100 | −0.094* | −0.174 | −0.015 |
| ZPP | Previous Hb | −0.010* | −0.020 | −0.001 | 0.003 | −0.008 | 0.013 |
| | Previous ZPP | 0.213* | 0.175 | 0.252 | 0.242* | 0.203 | 0.282 |

The asterisk (*) denotes that zero is not included in the 95% highest posterior density interval.

Our results suggest that the unobserved heterogeneity, which is expressed through the random effects, has a mutual influence on both variables. When the average ZPP level of an individual is higher than the population average ZPP level, then generally the average Hb level of that individual is lower than the population average of Hb levels, and vice versa. This implies that, when using ZPP

as an explanatory variable in a model which predicts Hb, one can also opt for using a long term average of ZPP as a predictor. This has the advantage that ZPP does not necessarily need to be measured at every visit.

## 6.2 Prediction results

In this subsection we analyze the performance of our models with regards to the prediction of future Hb levels. We first evaluate the added value of ZPP as an explanatory variable, after which we examine the performance of the different predictions models we argued for in Section 5. Lastly, we evaluate the benefits of the additional set of blood levels. In order to obtain the reported performance measures, we randomly split the data in a train and a test set ten times. We compute the relevant performance measures for each of the random splits, and we report the average performance measures over the ten replications. We use each model to predict Hb levels as a numeric outcome, and we report the RMSE and MAE as our error measures. We compare the discriminative ability of each model by examining the AUC measure that is calculated based on the predicted Hb levels (as also explained in Section 3.4). Note that a higher AUC value indicates better performance, as opposed to the other two reported performance measures. We observe that the AUC measure is relatively unstable, i.e., it is very dependent on the specific splits of the train and test sets. This means that the accuracy of the discrimination of our used methods is highly dependent on the data that we use to train and evaluate our models. This finding can possibly be explained by the small number of deferrals in our data.

### 6.2.1 The benefits of using ZPP as a predictor

In order to test the usefulness of ZPP for the prediction of Hb levels in an out-of-sample context, we compare the performance of three different linear mixed-effect transition models. Apart from the model in which we use all explanatory variables from Table 1, we also fit a model without previous ZPP measurements. Thus, this model uses all explanatory variables from Table 1, except for previous ZPP. Additionally, we fit a model for which we also use all explanatory variables, but we now replace the previous ZPP value with a dynamic individual-specific average of ZPP. Note that the 'model with previous ZPP' is our primary model, with which we have obtained the parameter estimates as shown in Table 3. In order to avoid any look-ahead bias, we use only observations up to time $t$ to compute the average ZPP that is used to predict Hb at time $t+1$.

**Table 7:** Average out-of-sample prediction performance measures for three different linear mixed-effect transition models

| | **Females** | | | **Males** | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | AUC | MAE | RMSE | AUC |
| Model without ZPP | 0.399 | 0.506 | 0.671 | 0.416 | 0.529 | 0.651 |
| Model with previous ZPP | 0.398 | 0.504 | 0.679 | 0.415 | 0.529 | 0.658 |
| Model with average ZPP | 0.399 | 0.506 | 0.675 | 0.415 | 0.529 | 0.658 |

The results in Table 7 suggest that the three models all perform very similar in terms of the used performance measures. The most important implication of this finding is that the added value of ZPP as a predictor for Hb seems to be absent in an out-of-sample context. In general, caution is required in extrapolating the findings of in-sample analyses to out-of-sample contexts. While findings of Baart

et al. (2013) suggested that ZPP are useful as a predictor for future Hb levels, they solely focus on in-sample analyses. Even though our models indicate that ZPP levels are associated with Hb levels, this relationship seems of limited usefulness for predicting future Hb levels or determining eligibility for donation. This result can possibly be explained by (i) the relatively small effect-size of ZPP and (ii) our finding that ZPP appears to be mostly useful for explaining average Hb levels. As the random intercept is already able to capture between-individual variation of Hb levels, adding measurements of ZPP to a regression model may have little added value from a forecasting perspective.

### 6.2.2 Comparison of different methods for predicting longitudinal outcomes

In this subsection we compare the out-of-sample performance of the models that we argued for in Section 5 to the more established models in longitudinal forecasting, namely the mixed-effects (transition) models, which are described in Section 3.1. The performance measures for the standard linear mixed-effects model (without autoregressive term) are also computed, because this model is very frequently used in current biostatistical research and therefore serves as our benchmark. Furthermore, this model has also been used in the context of Hb predictions by Nasserinejad et al. (2013), so computing the relevant performance measures for this model facilitates comparison to earlier research. Note that for each of the models that we evaluate in this section, we use the same set of explanatory variables to predict Hb levels. These explanatory variables are specified in Table 1.

**Table 8:** Average out-of-sample prediction performance measures for different models

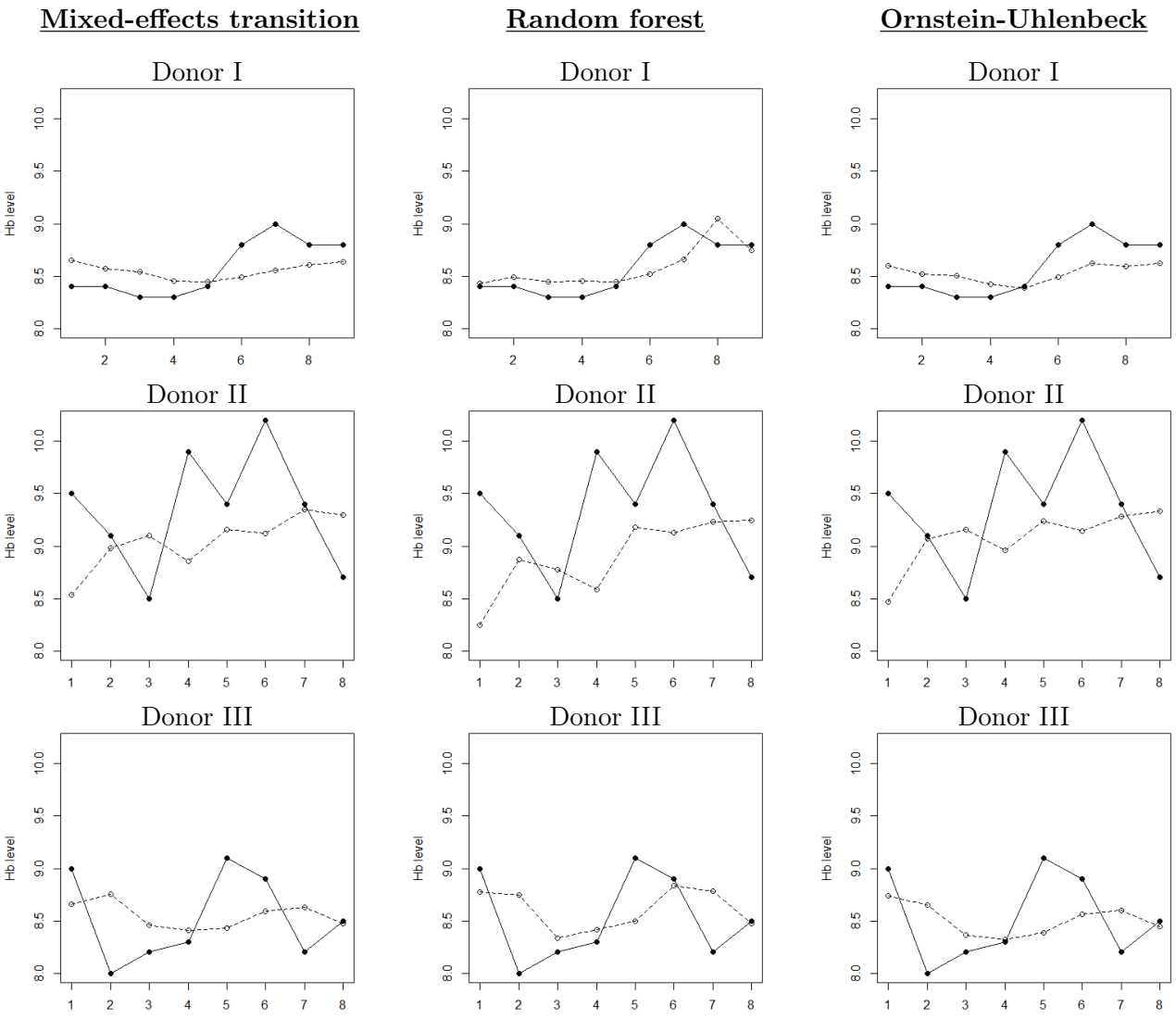
| | **Females** | | | **Males** | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | AUC | MAE | RMSE | AUC |
| Mixed-effects model | 0.411 | 0.512 | 0.661 | 0.421 | 0.536 | 0.648 |
| Mixed-effects transition model* | 0.398 | 0.504 | 0.679 | 0.415 | 0.529 | 0.658 |
| Random forest | 0.420 | 0.532 | 0.717 | 0.432 | 0.538 | 0.690 |
| Gradient tree boosting | 0.419 | 0.535 | 0.711 | 0.435 | 0.542 | 0.692 |
| Hierarchical Ornstein–Uhlenbeck | 0.389 | 0.499 | 0.694 | 0.409 | 0.514 | 0.666 |

*Note that this model refers to the standard mixed-effects transition model, which we also used to obtain the inferential results from Table 3. Hence, the reported performance measures are identical to those of the 'model with previous ZPP' from Table 7.

The results in Table 8 give various insights in the performance of the different models. The two decision tree ensemble methods perform best in terms of determining eligibility for donation, as they have the highest AUC values. Furthermore, the hierarchical Ornstein-Uhlenbeck process model performs the best in terms of RMSE and MAE, also outperforming the mixed-effects transition model. Overall, the performance of our models seems somewhat disappointing, especially when focusing on the AUC values. Visual inspection of the out-of-sample forecasts suggest that for a large number of donations, the predictions of the hierarchical Ornstein-Uhlenbeck are highly similar to the predictions of the autoregressive mixed-effects model, which can be also observed for the three randomly selected donors shown in Figure 3. The predictions of the Ornstein-Uhlenbeck model are on average slightly better, but the difference is limited. Plots of the predictions suggest that the traditional mixed-effects (transition) model and the hierarchical Ornstein-Uhlenbeck model are both successful in explaining the majority of the between-subject variation. The decision tree based methods are slightly less successful in doing so, which also explains why these models show higher

RMSE and MAE values. The inability of these models to accurately account for between-subject variation can be explained by the absence of explicit individual specific parameters in these models, which makes it not possible for the algorithms to account for heterogeneity that is not observed through the covariates in our data.

**Figure 3:** Predicted and observed Hb values for three randomly selected female donors in the test set. The black lines refer to the observed Hb values of each donors, and the grey lines refer to the corresponding predicted values.[8]




Overall, it seems that none of the methods are very well able to accurately explain deviations of the expected trajectories of Hb levels for any specific individual. When shocks occurs, i.e., when a measurement is substantially higher or lower than previous measurements, we would ideally be able to link this to external factors, such as the season in which the donation takes place or the ZPP

[8] We randomly selected three donors from the subset of donors in one of the test sets that had at least 5 donations.

measurement at previous visit. In practice, we are only able to do so on a limited scale, which is in line with our in-sample findings of Section 6.1 that suggested that the explanatory power of the covariates in our data is in general very low. The decision tree based methods seem slightly more successful in exploiting the explanatory variables than the other methods, resulting in a relatively high discriminative ability of the decision tree models. This implies nonlinear relationships might be relevant for explaining the Hb trajectories. The somewhat disappointing accuracy of our predictions is likely to be (partially) blameable on our data: in Nasserinejad et al. (2013) the same implementation of the mixed-effects model achieved an out-of-sample AUC of 0.81 for men, while for us this model only scores an AUC of 0.65 (and is our worst performing model). Possible reasons for this difference are (i) the availability of different explanatory variables, (ii) a larger sample size, and in particular (iii) a generally larger number of observations for each individual. In our data we are dealing with a relatively large number of donors who only donate less than three times during the observational period, and the Hb levels for these individuals are in general harder to predict. This notion is confirmed when we compare the predictive performance of our models across individuals with differing numbers of donations. Especially for the mixed-effects regression models and the Ornstein-Uhlenbeck model we find that individuals with a higher number of visits during our observational period typically yield better out-of-sample performance measures.

Further inspection of plots of observed Hb level trajectories of various individuals indicates that the theoretical trajectory, which implies that Hb levels are lowered by donations and afterwards slowly recover to their 'normal levels' after each donation, cannot be clearly observed in our data. This unexpected data characteristic, along with the absence of a larger explanatory power of our predictor variables, might also be partially caused by the noisiness of the measurements of Hb. Every measurement of Hb is obtained by means of a finger-stick sample, which can lead to relatively inadequate measurements. When the Hb level of an individual is measured twice, the two measurement can in practice be quite different, even if the second measurement occurred immediately after the first.

### 6.2.3 The added value of additional blood levels as explanatory variables

In order to assess the added value of the additional blood levels when it comes to predicting Hb levels, we compare the performance of the mixed-effects transition model in which we incorporate the additional set of blood levels by means of the spike and slab regression described in Section 5.3 to our standard mixed-effects transition model, without including the extra variables. We estimate and evaluate both models only on the subset of donors for which the additional set of blood levels is available. We consider two alternative specifications of the spike and slab regression model. In the first one we restrict the more established predictors, which are also used to gather the inferential results in for example Table 3, to be included in the model. For the second one, we only restrict the intercept and the previous Hb values to be included in the model, and therefore we are applying variable selection on all the other explanatory variables.

In the subset of the data for which the additional blood levels are available, the number of deferrals is very low; for a substantial number of train-test splits there was not even a single deferral included in the test set. In order to still be able to evaluate the discriminative ability of our models, we redefine the AUC measure: we will now investigate how accurate our models can predict donations to be lower than the thresholds 7.9 and 8.6, which correspond to the lowest 10% percentile

measurements of females and males, respectively. In order to improve model fit, we used a log transformation on some of the blood levels that had skewed data distributions.

**Table 9:** Average out-of-sample prediction performance measures for the subset of the data for which the additional blood levels are available

| | **Females** | | | **Males** | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | AUC | MAE | RMSE | AUC |
| Mixed-effects transition | 0.398 | 0.508 | 0.681 | 0.413 | 0.526 | 0.722 |
| Mixed-effects transition + blood levels (**i**) | 0.389 | 0.499 | 0.696 | 0.401 | 0.512 | 0.748 |
| Mixed-effects transition + blood levels (**ii**) | 0.386 | 0.496 | 0.696 | 0.398 | 0.510 | 0.748 |

For model (**i**) we restrict the variables as shown in Table 3 to be included in the model at each iteration. For model (**ii**) we only restrict the intercept and the previous Hb levels to be always included in the model.

The results in Table 9 show that incorporating the additional blood levels leads to a moderate increase of the average performance. That is, the use of the additional variables results in small improvements of the out-of-sample prediction accuracy measures in the range of 1 to 4 percent. From a practical perspective it is very questionable whether such limited improvements of the out-of-sample accuracy are sufficient to motivate the actual use of these variables in clinical practice. As the used blood levels cannot be obtained with a standard fingerstick capillary sample, it is more costly to incorporate measurements of these variables in a donation procedure.

It might be useful to select only a limited number of variables for actual use, where the selection can be based on the in-sample inclusion probabilities. For example, the red blood cell count (RBC) and hematocrit (HCT) consistently rank high in terms of inclusion probability, which indicates these variables play an important role in reducing the in-sample prediction errors. The results in Table 9 do again not answer any questions on the exact relation of the various blood characteristics and Hb levels. It could be the case that it is mostly the correlation of average values of Hb and the various other blood levels that drives the improvement of the predictions, rather than that the blood levels at time $t-1$ can really explain Hb levels at time $t$. Further research is needed to examine these relations, but our results suggest that these variable may potentially be useful for the prediction of future Hb levels.

# 7 Conclusion, discussion

In this thesis we have considered various statistical techniques to (i) examine the longitudinal association of ZPP and Hb and (ii) predict future Hb levels. For the purpose of forecasting future Hb levels we utilized three different types of models: an autoregressive random intercept model, two decision tree based ensemble methods and a hierarchical specification of the Ornstein-Uhlenbeck model. Additionally, we used spike and slab regression to incorporate an relatively high-dimensional set of blood levels into our models. In order to investigate the exact association of Hb and ZPP, we employed a multivariate autoregressive mixed-effects model.

With the application of our multivariate autoregressive mixed-effects model we illustrate that there is a subtle difference between state dependence and unobserved heterogeneity in a multivariate context. We also show how inaccurate model specifications can have a detrimental effect on the discovery of the true association. Previous econometric literature such as Heckman (1981) and Keane (1997) already defined this difference between unobserved heterogeneity and state dependence, and we are able to illustrate this difference in a multivariate context by means of a practical example. When simultaneously modelling unobserved heterogeneity and state dependence, the traditional model assumptions are violated. In a simulation experiment we empirically verify the validity of multivariate generalizations of existing methods to correct for the violation of these model assumption.

We find that our data indicates that the association between ZPP and Hb is not reflected through the cross-lagged effects of the two variables, but rather through the negative correlation of the random intercepts of individuals. From a clinical perspective this finding can be reformulated by stating that the unobserved heterogeneity of individuals has a opposite influence on the two biomarkers. That is, unobserved factors like genetics, diet and exercise, which are not measured in our data, have a concurrent influence on ZPP and Hb. Therefore, the average value of one of the two variables for a specific individual provides some indication for the average value of the other variable. It may be feasible to use an average of ZPP as a predictor of Hb values in order to schedule future donation moments. Yet, our further analyses suggested that the usefulness of ZPP for the (out-of-sample) prediction of future Hb levels is very limited. Therefore, we are not able to generalize Baart et al.'s (2013) in-sample findings to an out-of-sample context, and it remains very questionable whether ZPP should be used to predict Hb levels in a practical setting.

When comparing the predictive performance of our different models, we find that the proposed hierachical Ornstein-Uhlenbeck process model on average outperforms traditional mixed-effects models. Additionally, we find that incorporating blood levels can improve Hb predictions and that the decision tree ensemble methods are most successful in determining eligibility of donation. However, despite the use of different types of state-of-the-art prediction methods, the out-of-sample results are not fully satisfying. Especially in terms of discriminative ability, it seems that none of the models is very successful in identifying whether donors will be eligible for donation. This can mostly be blamed on the limited usefulness of the set of explanatory variables, which seems unable to accurately explain the within-subject variation that is present in our data. This notion is supported by our in-sample analyses, which suggest that the explanatory power of our set of predictor variables is in general quite low.

It may be a future aim to combine the properties of the different models. While the mixed-effects models and the hierachical Ornstein-Uhlenbeck process model were typically successful in explaining the majority of the between-subject variation, they failed to accurately capture the within-subject variation. Conversely, the two decision tree based methods were better able to explain deviations of the expected Hb trajectories of individuals, but failed to accurately capture individual heterogeneity that is not revealed through the covariates in our data. The most straightforward way to combine the different properties of our models can be achieved by using model stacking, which is especially popular outside academia. From a practical perspective, it may also be useful to evaluate other (machine learning) models that are able to account for nonlinear relationships of our variables. The most obvious extension would be to test predictions that are based on neural networks.

We have several more suggestions for future research. From a methodological perspective, it could be very interesting to combine the Ornstein-Uhlenbeck model specification with spike and slab variable selection on the explanatory variables of our model. That is, one could consider a spike and slab prior on the parameters related to the external explanatory variables in Equation (36). The resulting model could be used for a variety of (biomedical) applications, where a biomarker has a stationary trajectory and is possibly related to a (large) set of external predictors. The idea behind this model is similar to the idea of the BSTS model of Scott and Varian (2014), which combines spike and slab regression with a Kalman Filter and has proven to be successful in i.a. forecasting macroeconomic time series using external predictors. Whilst in our case the relatively moderate number of explanatory variables might not make the variable selection an essential property, higher dimensional biomedical regression problems, such as problems related to gene expression data, could very well profit from incorporating the spike and slab prior.

For inferential purposes, the hierarchical Ornstein-Uhlenbeck process model could possibly also help to discover the recovery time of hemoglobin that is needed between two subsequent donations. In Nasserinejad et al. (2016) a function of various parameters was added to a linear regression model for the same purpose. The decay parameter in the Ornstein-Uhlenbeck model can be interpreted as the speed to which the stochastic process returns to its asymptotic mean. Therefore, it could possibly explain how quickly Hb levels return to their individual specific mean, and thus how quickly an individual recovers from donating blood. It could also be beneficial to incorporate additional parameters that specifically capture the impact of a donation, similar to what was done in Nasserinejad et al. (2016). Additionally, it can be useful to incorporate a method to explicitly deal with measurement errors, as these are not automatically accounted for by the Ornstein-Uhlenbeck model.

## References


Akay, A. (2012). Finite-sample comparison of alternative methods for estimating dynamic panel data models. *Journal of Applied Econometrics*, 27(7):1189–1204.

Arulampalam, W. and Stewart, M. B. (2009). Simplified implementation of the heckman estimator of the dynamic probit model and a comparison with alternative estimators. *Oxford bulletin of economics and statistics*, 71(5):659–681.

Baart, A. M., Kort, W. L., Moons, K. G., Atsma, F., and Vergouwe, Y. (2013). Zinc protoporphyrin levels have added value in the prediction of low hemoglobin deferral in whole blood donors. *Transfusion*, 53(8):1661–1669.

Boulton, F. (2004). Managing donors and iron deficiency. *Vox sanguinis*, 87:22–24.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees, the wadsworth statistics and probability series, wadsworth international group, belmont california (pp. 356).

Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9):2812–2831.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

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

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Crowell, R., Ferris, A. M., Wood, R. J., Joyce, P., and Slivka, H. (2006). Comparative effectiveness of zinc protoporphyrin and hemoglobin concentrations in identifying iron deficiency in a group of low-income, preschool-aged children: practical implications of recent illness. *Pediatrics*, 118(1):224–232.

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

Fok, D., Horváth, C., Paap, R., and Franses, P. H. (2006). A hierarchical bayes error correction model to explain dynamic effects of price changes. *Journal of Marketing Research*, 43(3):443–461.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Greenberg, E. (2012). *Introduction to Bayesian econometrics*. Cambridge University Press.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Heckman, J. J. (1981). Heterogeneity and state dependence. In *Studies in labor markets*, pages 91–140. University of Chicago Press.

Heckman, J. J. (1987). *The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some Monte Carlo evidence*. University of Chicago Center for Mathematical studies in Business and Economics.

Heij, C., de Boer, P., Franses, P. H., Kloek, T., van Dijk, H. K., et al. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

Hoekstra, T., Veldhuizen, I., Van Noord, P., and De Kort, W. (2007). Seasonal influences on hemoglobin levels and deferral rates in whole-blood and plasma donors. *Transfusion*, 47(5):895–900.

Hsiao, C. (2014). *Analysis of panel data*. Number 54. Cambridge university press.

Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Janz, T. G., Johnson, R. L., and Rubenstein, S. D. (2013). Anemia in the emergency department: evaluation and treatment. *Emergency medicine practice*, 15(11):1–15.

Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer.

Kazemi, I. and Crouchley, R. (2006). Modelling the initial conditions in dynamic regression models of panel data with random effects. *Contributions to Economic Analysis*, 274:91–117.

Keane, M. P. (1997). Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, 15(3):310–327.

Kiss, J. E., Brambilla, D., Glynn, S. A., Mast, A. E., Spencer, B. R., Stone, M., Kleinman, S. H., and Cable, R. G. (2015). Oral iron supplementation after blood donation: a randomized clinical trial. *Jama*, 313(6):575–583.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.

Martin, C., Werntz III, C., and Ducatman, A. (2004). The interpretation of zinc protoporphyrin changes in lead intoxication: a case report and review of the literature. *Occupational Medicine*, 54(8):587–591.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Nadler, S. B., Hidalgo, J. U., and Bloch, T. (1962). Prediction of blood volume in normal human adults. *Surgery*, 51(2):224–232.

Nasserinejad, K., de Kort, W., Baart, M., Komárek, A., van Rosmalen, J., and Lesaffre, E. (2013). Predicting hemoglobin levels in whole blood donors using transition models and mixed effects models. *BMC medical research methodology*, 13(1):62.

Nasserinejad, K., Rosmalen, J. v., Kort, W. d., Rizopoulos, D., and Lesaffre, E. (2016). Prediction of hemoglobin in blood donors using a latent class mixed-effects transition model. *Statistics in medicine*, 35(4):581–594.

Orme, C. D. (1996). *The initial conditions problem and two-step estimation in discrete panel data models*. University of Manchester.

Oskrochi, G., Lesaffre, E., Oskrochi, Y., and Shamley, D. (2016). An application of the multivariate linear mixed model to the analysis of shoulder complexity in breast cancer patients. *International journal of environmental research and public health*, 13(3):274.

Plummer, M. (2012). Jags version 3.3. 0 user manual. *International Agency for Research on Cancer, Lyon, France*.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scott, S. L. and Varian, H. R. (2014). Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23.

Simon, T. L. (2002). Iron, iron everywhere but not enough to donate. *Transfusion*, 42(6):664–664.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stolwijk, A., Straatman, H., and Zielhuis, G. (1999). Studying seasonality by using sine and cosine functions in regression analysis. *Journal of Epidemiology & Community Health*, 53(4):235–238.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Introduction to data mining. 1st.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2):177–188.

Verbeke, G. (1997). Linear mixed models for longitudinal data. In *Linear mixed models in practice*, pages 63–153. Springer.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research*, 23(1):42–59.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of applied econometrics*, 20(1):39–54.

Yip, R., Johnson, C., and Dallman, P. R. (1984). Age-related changes in laboratory values used in the diagnosis of anemia and iron deficiency. *The American journal of clinical nutrition*, 39(3):427–436.

# Appendices

## A Initial conditions problem in a multivariate setting: a simulation study

Even though the initial conditions problem that arises when using autoregressive mixed-effect models is well covered in statistical literature, the validity of the established solutions to this problem in a multivariate setting are not yet empirically examined. Multivariate autoregressive mixed-effects models have an increased complexity, due to the allowance of correlation of the residuals and the individual specific intercepts. Furthermore, previous simulation experiments that evaluated solutions to the initial conditions problem considered only very low dimensional regression designs, which results in limited generalizability of the obtained findings. In this section we provide a small simulation study to analyze the performance of the Heckman solution and the Wooldridge solution in a multivariate regression model, and we compare their performance to an analysis where the initial conditions problem is neglected. We also consider a more extensive data generating process than in previous simulation studies.

### A.1 Simulation design

With the set-up of the simulation study we aim to generate data in a way that closely resembles a data generating process (DGP) of real longitudinal outcomes, such as biomarkers (like ZPP and Hb). The simulation set-up is relatively complicated because we need to take into account (i) the possible association structures of the two response variables, (ii) the longitudinal structure of the data and corresponding structure of the predictor variables and (iii) the notion that it is not reasonable to assume that the DGP starts at the first observation. In order to evaluate the performance of the three different models, we generate 100 artificial datasets. The procedure that we use to generate these datasets is explained in the next subsections.

#### A.1.1 General data generating process

In order to simulate data that has the appropriate longitudinal structure, we consider a case where we have repeated observations of $n = 300$ individuals. For these $n$ individuals we consider three balanced panel situations where each individual has either $t = 5$, 10 or 15 measurements, which leads to total samples sizes of 1500, 3000 and 4500 observations. We also consider a situation where we are dealing with an unbalanced panel situation, in which we generate a varying number of observations for every individual. We generate this number of observations by drawing from the number of donations for each individual that we find in our real data, described in Section 2. As a consequence, the structure of our artificial unbalanced panel datasets should be very comparable to the structure that we find in the Sanquin data.

We consider the following data generating process (DGP) to simulate the response variables:

$$\begin{pmatrix} y_{1,it} \\ y_{2,it} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} b_{1,i} \\ b_{2,i} \end{pmatrix} + \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix} \begin{pmatrix} y_{1,i,t-1} \\ y_{2,i,t-1} \end{pmatrix} + \begin{pmatrix} \mathbf{x}'_{it}\boldsymbol{\beta}_1 \\ \mathbf{x}'_{it}\boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,it} \\ \varepsilon_{2,it} \end{pmatrix}, \tag{42}$$

$$\boldsymbol{\varepsilon}_{it} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}), \quad \boldsymbol{\alpha} = (8, 5)', \quad \boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}). \tag{43}$$

The terms $\boldsymbol{\Sigma_b}, \boldsymbol{\Sigma_\varepsilon}$, $\boldsymbol{A}$ and $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ are the model parameters that define the association structure of the two response variables. The matrix $\boldsymbol{X}$ consists of 9 explanatory variables, whose simulation we describe in Section A.1.2.

The first observation of the two response variables of an individual is generated by means of:

$$\boldsymbol{y}_{i0} = \begin{pmatrix} \boldsymbol{x}'_{i0}\boldsymbol{\beta}_1 \\ \boldsymbol{x}'_{i0}\boldsymbol{\beta}_2 \end{pmatrix} + 1.1\boldsymbol{\alpha} + \boldsymbol{b}_i + \boldsymbol{\varepsilon}_{i0}. \tag{44}$$

Subsequent observations are generated by means of the DGP defined in Equation (42), so with the values of the autoregressive term included in the equation. In general, it is not reasonable to assume the process that is analyzed in longitudinal studies starts at the first measurement. For example, the first measurement of Hb at the first donation of a specific individual does not correspond to the first realization of this stochastic variable; each human being has had Hb levels fluctuating since they were born. In order to incorporate this notion into our simulation study, we generate observations prior to the start of the observational period (the set of observations that are used to estimate the models). Therefore, we generate 10 extra observations for each individual in all of our datasets. For example, for our balanced panel design of $t = 5$, we generate observations for $t = 0, 1, ..., 15$. When estimating the parameters of our models, the first 10 observations of each individual are discarded, so that we are left with 5 observations for each individual.

The full set of parameters that we use to generate our response variable is defined as follows:

$$\boldsymbol{\Pi} = \begin{pmatrix} \boldsymbol{\beta}'_1 \\ \boldsymbol{\beta}'_2 \end{pmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{19} \\ \beta_{21} & \beta_{22} & \dots & \beta_{29} \end{bmatrix}. \tag{45}$$

We generate a new parameter matrix $\boldsymbol{\Pi}$ for every simulated dataset. In order to do so, we simulate each of the $2 \times 9$ elements in $\boldsymbol{\Pi}$ separately by setting $\beta_{ij} = C$. Here $C$ is a discrete random variable, with $p(C = 0.5) = 0.25$, $p(C = -0.5) = 0.25$ and $p(C = 0.0) = 0.5$. Thus, for every value of $\beta_{ij}$, we simulate a separate value $C$. The random variable $C$ is independently and identically distributed.

### A.1.2 Simulation of predictor variables

In longitudinal studies it is reasonable to assume that the predictor variables are composed of a mix different types of variables. We define three types of variables, and consider the notation as specified below.

- $\check{\boldsymbol{x}}$: variables that are constant for a specific individual, such as menopausal status,
- $\ddot{\boldsymbol{x}}$: variables that only rely on timing of the measurement, such as seasonal variables,
- $\tilde{\boldsymbol{x}}$: variables that are individual specific, and do change over time, such as blood pressure.

We simulate the initial values for $\check{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}$ for every $i$ by means of:

$$\begin{aligned} &(\check{\boldsymbol{x}}_{i0}, \tilde{\boldsymbol{x}}_{i0}) \sim N(\boldsymbol{\mu}_{1:6}, \boldsymbol{\Sigma}_{\boldsymbol{x}0}) \\ &\boldsymbol{\mu} = (1, 0, 1, 1, 0, 1, 1, 0, 1)' \end{aligned} \tag{46}$$

In order to create predictor variables that are correlated, we set $\boldsymbol{\Sigma}_{\boldsymbol{x}0}(i, i) = 1$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}0}(i, j) = 0.5^{|i-j|}\ \forall\ i \neq j$. The covariates in $\ddot{\boldsymbol{x}}$ are assumed to be independent of the other covariates and are generated by means of:

$$\ddot{\boldsymbol{x}}_{i0} \sim N(\boldsymbol{\mu}_{7:9}, \mathbf{I}_3). \tag{47}$$

Given the initial values of the three types of predictors for an individual $i$, we generate subsequent values by means of:

$$\ddot{\boldsymbol{x}}_{it} \sim N(\boldsymbol{\mu}_{7:9}, \mathbf{I}_3), \quad \check{\boldsymbol{x}}_{it} = \check{\boldsymbol{x}}_{it-1}, \quad \tilde{\boldsymbol{x}}_{it} = \tilde{\boldsymbol{x}}_{i0} + \boldsymbol{\xi}_{it} \text{ with } \boldsymbol{\xi}_{it} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}}). \tag{48}$$

The covariance matrix of the disturbances in $\boldsymbol{\xi}_{it}$ is defined as $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}(i,i) = 0.2$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}(i,j) = 0.1^{|i-j|} \; \forall \, i \neq j$, so that the changes of the variables in $\tilde{\boldsymbol{x}}_{it}$ have a relatively strong correlation.

### A.1.3 Association structure scenarios

The parameter values that are generated in $\boldsymbol{\Pi}$ already create correlation between the two response variables, as the mutual non-zero elements in $\boldsymbol{\Pi}$ result in an association of the two variables. Additionally, we consider the following three scenarios for the covariance matrix of the residuals, the covariance matrix of the random effects and the autoregressive matrix:

- (**1**) *Only cross-lagged effect.*

$$\boldsymbol{\Sigma}_{\boldsymbol{b}} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}, \qquad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \tag{49}$$

- (**2**) *Full dependence structure.*

$$\boldsymbol{\Sigma}_{\boldsymbol{b}} = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}, \qquad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix} \tag{50}$$

- (**3**) *Full dependence structure with increased heterogeneity.*

$$\boldsymbol{\Sigma}_{\boldsymbol{b}} = \begin{pmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{pmatrix}, \qquad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix} \tag{51}$$

In situation (**1**), the only association of the two response variables arises because of the parameter values in $\boldsymbol{\Pi}$ and because of the off-diagonal elements in $\boldsymbol{A}$. In situation (**2**), we consider a situation where the association arises because of the off-diagonal in $\boldsymbol{A}$, but also due to the correlation of the residuals and the random effects. We let the correlation of the random effects be higher than the correlation of the residuals, as this is in line what we find in our real data. In situation (**3**), we consider a situation comparable to (**2**), but now with larger variances (and covariances) of the random effects. Therefore, in situation (**3**) we increased the presence of the heterogeneity amongst individuals.

### A.1.4 Performance evaluation

In order to evaluate the estimation performance of the different methods, we will consider the root mean square error (RMSE) and the average error (bias) of the estimated parameters with regards to the true DGP parameters over the 100 generated datasets. We consider the errors for the coefficients in $\boldsymbol{\Pi}$, the diagonal elements of $\boldsymbol{A}$ and the off-diagonal elements of $\boldsymbol{A}$ separately. Thus, we define three sets of parameters: $\boldsymbol{\theta}_{\boldsymbol{A}_1}$, $\boldsymbol{\theta}_{\boldsymbol{A}_2}$ and $\boldsymbol{\theta}_{\boldsymbol{\Pi}}$, with $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ referring to the diagonal and off-diagonal elements in $\boldsymbol{A}$, respectively. For each set of parameters, the performance measures are given by:

$$\text{Bias}(\hat{\boldsymbol{\theta}}) = \mathbb{E}(\hat{\theta} - \theta) \tag{52}$$

$$\text{RMSE}(\hat{\boldsymbol{\theta}}) = \left[\mathbb{E}(\hat{\theta} - \theta)^2\right]^{\frac{1}{2}}, \tag{53}$$

where the term $\hat{\theta}$ refers to the posterior mean of each estimated parameter. In order to obtain the expected value for both error measures, we consider the average over all 100 replications and over all parameters in $\boldsymbol{\theta}$. The RMSE can be decomposed as the square root of the sum of the variance of the estimator and the squared bias of the estimator. Therefore, by evaluating both the bias and RMSE, we can also say something about the variance of the estimator, as this is the second component that is reflected in the RMSE.

## A.2 Results of the simulation study

**Table 10:** Unbalanced panel results under situation (**1**): *Only cross-lagged effect*

| | **RMSE** | | | **Bias** | | |
|---|---|---|---|---|---|---|
| | No correction | Heckman | Wooldridge | No correction | Heckman | Wooldridge |
| full parameter matrix $\boldsymbol{\Pi}$ | 0.065 | 0.072 | 0.094 | -0.001 | 0.000 | -0.004 |
| diagonal elements $\boldsymbol{A}$ | 0.134 | 0.082 | 0.029 | 0.102 | 0.008 | 0.001 |
| off-diagonal elements $\boldsymbol{A}$ | 0.050 | 0.034 | 0.036 | 0.023 | -0.018 | -0.014 |

**Table 11:** Unbalanced panel results under situation (**2**): *Full dependence structure*

| | **RMSE** | | | **Bias** | | |
|---|---|---|---|---|---|---|
| | No correction | Heckman | Wooldridge | No correction | Heckman | Wooldridge |
| full parameter matrix $\boldsymbol{\Pi}$ | 0.080 | 0.077 | 0.110 | -0.003 | -0.000 | -0.005 |
| diagonal elements $\boldsymbol{A}$ | 0.153 | 0.101 | 0.038 | 0.124 | 0.010 | 0.002 |
| off-diagonal elements $\boldsymbol{A}$ | 0.103 | 0.049 | 0.057 | 0.076 | 0.019 | 0.042 |

**Table 12:** Unbalanced panel results under situation (**3**): *Full dependence structure with increased heterogeneity.*

| | **RMSE** | | | **Bias** | | |
|---|---|---|---|---|---|---|
| | No correction | Heckman | Wooldridge | No correction | Heckman | Wooldridge |
| full parameter matrix $\boldsymbol{\Pi}$ | 0.110 | 0.097 | 0.127 | -0.009 | -0.002 | -0.009 |
| diagonal elements $\boldsymbol{A}$ | 0.215 | 0.131 | 0.045 | 0.179 | 0.017 | 0.002 |
| off-diagonal elements $\boldsymbol{A}$ | 0.155 | 0.065 | 0.059 | 0.117 | 0.017 | 0.037 |

Tables 10, 11 and 12 show the average performance of the three methods for the unbalanced panel data under the three different association structure scenarios. The results confirm the classical econometrical claim that ignoring the initial conditions problem leads to an upward bias of the autoregressive parameters. That is, the estimates for the diagonal elements of $\boldsymbol{A}$ show a substantial bias. Even in a situation with only cross-lagged effects, for which the results are shown in Table 10, the percentile bias of the autoregressive parameter is more than 33%, while for the other situations the bias is even worse. The Heckman and the Wooldridge solution seem both well able to correct for this bias. The Wooldridge solution appears to be slightly more successful in doing so and approaches a bias of zero. The RMSE values are quite still far from zero, with especially the Heckman implementation having a substantial RMSE value for the diagonal elements of $\boldsymbol{A}$. As the RMSE

reflects both the bias and the variance of an estimator, the results suggest that the Heckman and Wooldridge estimates have a slightly lower efficiency than the estimates of the normal 'no correction' model, which neglects the initial conditions problem.

When comparing the unbalanced panel results of the three different association structure scenarios, we find that an increased complexity of the association structure worsens the bias of the estimates of the 'no correction' model. When there is correlation of the random intercepts, which reflects the unobserved heterogeneity, the upward bias for the state dependence parameters increases. Furthermore, when the severity of heterogeneity among individuals increases, the bias also gets considerably worse. Therefore, our results suggests that the higher the complexity of the association and the higher the degree of unobserved heterogeneity, the more desirable it is to account for the initial conditions problem by means of a one of the two evaluated solutions.

The initial conditions problem does not seem to have a severe impact on the estimate accuracy of the parameter matrix $\mathbf{\Pi}$. That is, the downward bias of the estimated coefficients of the external explanatory variables, which is described in Kazemi and Crouchley (2006), seems very limited in practice. Looking at the results for the off-diagonal elements of $\boldsymbol{A}$, which define the cross-lagged effects, we find several interesting results. The previous measurements of the other response variable behave very different than the other explanatory variables, as the corresponding parameters show a considerably larger bias. Furthermore, the Heckman solution appears to be more successful than the Wooldridge implementation in solving this bias. Even though the Heckman solution performs the best, the estimates of the corresponding model can still considered to be biased. Note that for the DGP parameters of the off-diagonal elements of $\boldsymbol{A}$ we used the value 0.1, so the bias of 0.017 of the Heckman implementation corresponds to an upward bias of 17%.

**Table 13:** Balanced panel results under situation (**2**) *Full dependence structure.*

| | | **RMSE** | | | **Bias** | | |
|---|---|---|---|---|---|---|---|
| | | No correction | Heckman | Wooldridge | No correction | Heckman | Wooldridge |
| $t = 5$ | full parameter matrix $\mathbf{\Pi}$ | 0.107 | 0.092 | 0.115 | -0.004 | -0.001 | -0.005 |
| | diagonal elements $\boldsymbol{A}$ | 0.210 | 0.130 | 0.070 | 0.191 | 0.020 | 0.004 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.137 | 0.061 | 0.078 | 0.114 | 0.038 | 0.061 |
| $t = 10$ | full parameter matrix $\mathbf{\Pi}$ | 0.036 | 0.050 | 0.085 | -0.002 | -0.002 | -0.003 |
| | diagonal elements $\boldsymbol{A}$ | 0.043 | 0.043 | 0.020 | 0.021 | 0.001 | 0.000 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.033 | 0.024 | 0.031 | 0.010 | -0.003 | 0.019 |
| $t = 15$ | full parameter matrix $\mathbf{\Pi}$ | 0.028 | 0.038 | 0.062 | -0.002 | -0.001 | -0.004 |
| | diagonal elements $\boldsymbol{A}$ | 0.013 | 0.012 | 0.013 | 0.005 | 0.000 | 0.000 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.011 | 0.011 | 0.017 | 0.002 | -0.002 | 0.009 |

The results in Table 15 suggest that the severity of the bias of the diagonal elements of $\boldsymbol{A}$ decreases as the panel length increases. This is line with econometrical literature on dynamic panel models (e.g. Cameron and Trivedi, 2005). Again, the bias of the parameters in $\mathbf{\Pi}$ can be considered neglectable, even for shorter panel lengths. The bias of the estimates of the off-diagonal elements in $\boldsymbol{A}$ is most severe (when expressed in percentages). For shorter panel lengths the Wooldridge and the Heckman solution seem to be able to correct for some of this bias, but the estimates are still on average quite a bit too high compared to the DGP values. Even with a panel length of $t = 15$, the bias of 9% for Wooldridge estimate of these parameters remains remarkable. Interestingly, this bias

of the Wooldridge estimates is also higher than the bias of the 'no correction' model.

For the balanced panel data under the other scenarios, we observe similar patterns as described earlier. That is, the results show that an increasing panel length decreases the bias of the 'no correction' model estimates. Yet, under the increased heterogeneity scenario, the results are more in favor of the two models which do actually implement a correction for the initial conditions problem. Even for a panel length of $t = 15$, the bias of the estimates of the diagonal elements in $\boldsymbol{A}$ for the 'no correction' model is equal to 0.014, which corresponds to an upward bias of around 5%.

While in general the results of our simulations yield clear conclusions, it might be desirable to extend the simulations to more different situations. Considering all possible choices for the covariance matrices, the random effects, the matrix $\boldsymbol{A}$, the panel length and the set-up of the simulation regarding the simulation of the explanatory variables leaves us with an immense number of possible simulation scenarios. It warrants further research to specifically addresses the question which estimation method works best in what scenario.

**Table 14:** Balanced panel results under situation (**1**) *Only cross-lagged effect*

| | | **RMSE** | | | **Bias** | | |
|---|---|---|---|---|---|---|---|
| | | No correction | Heckman | Wooldridge | No correction | Heckman | Wooldridge |
| $t = 5$ | full parameter matrix $\boldsymbol{\Pi}$ | 0.109 | 0.098 | 0.120 | 0.003 | 0.003 | 0.004 |
| | diagonal elements $\boldsymbol{A}$ | 0.227 | 0.152 | 0.039 | 0.200 | 0.023 | 0.002 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.106 | 0.039 | 0.041 | 0.056 | -0.025 | -0.003 |
| $t = 10$ | full parameter matrix $\boldsymbol{\Pi}$ | 0.033 | 0.051 | 0.085 | -0.001 | -0.001 | -0.005 |
| | diagonal elements $\boldsymbol{A}$ | 0.027 | 0.039 | 0.018 | 0.020 | 0.001 | 0.001 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.018 | 0.018 | 0.022 | 0.009 | -0.006 | 0.004 |
| $t = 15$ | full parameter matrix $\boldsymbol{\Pi}$ | 0.028 | 0.037 | 0.065 | 0.000 | 0.000 | -0.005 |
| | diagonal elements $\boldsymbol{A}$ | 0.013 | 0.010 | 0.011 | 0.005 | 0.000 | 0.000 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.012 | 0.011 | 0.017 | 0.002 | -0.003 | 0.000 |

**Table 15:** Balanced panel results under situation (**3**) *Full dependence structure with increased heterogeneity*

| | | **RMSE** | | | **Bias** | | |
|---|---|---|---|---|---|---|---|
| | | No correction | Heckman | Wooldridge | No correction | Heckman | Wooldridge |
| $t = 5$ | full parameter matrix $\boldsymbol{\Pi}$ | 0.149 | 0.126 | 0.141 | 0.006 | 0.002 | 0.007 |
| | diagonal elements $\boldsymbol{A}$ | 0.291 | 0.184 | 0.065 | 0.266 | 0.042 | 0.004 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.172 | 0.077 | 0.076 | 0.147 | 0.026 | 0.051 |
| $t = 10$ | full parameter matrix $\boldsymbol{\Pi}$ | 0.086 | 0.081 | 0.116 | -0.006 | -0.004 | -0.002 |
| | diagonal elements $\boldsymbol{A}$ | 0.149 | 0.083 | 0.030 | 0.082 | 0.006 | 0.001 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.119 | 0.046 | 0.049 | 0.064 | 0.004 | 0.030 |
| $t = 15$ | full parameter matrix $\boldsymbol{\Pi}$ | 0.043 | 0.055 | 0.102 | -0.002 | 0.001 | -0.005 |
| | diagonal elements $\boldsymbol{A}$ | 0.043 | 0.057 | 0.013 | 0.014 | 0.004 | 0.000 |
| | off-diagonal elements $\boldsymbol{A}$ | 0.031 | 0.024 | 0.022 | 0.005 | -0.002 | 0.013 |

# B Full conditional posteriors of the (autoregressive) mixed-effects model

**Sampling of $\sigma_\varepsilon^2$**
The full conditional posterior of $\sigma_\varepsilon^2$ is given by:

$$p(\sigma_\varepsilon^2|\cdot) \propto \sigma_\varepsilon^{(\iota+\nu_\epsilon-1)/2}\exp\Big[-\frac{1}{2}\sigma_\varepsilon^{-2}(\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-b_i)'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-b_i)+ss_\varepsilon)\Big], \tag{54}$$

where $\iota$ equals the total number of observations in our data ($\sum_{i=1}^{n} T_i$). From the conditional posterior it follows that we can sample $\sigma_\varepsilon^2$ from an inverse gamma distribution with scale parameter $(\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-b_i)'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-b_i)+ss_\varepsilon)/2$ and degrees of freedom equal to $(\iota+\nu_\varepsilon)/2$.

**Sampling of $\sigma_b^2$**
The full conditional posterior of $\sigma_b^2$ is given by:

$$p(\sigma_b^2|\cdot) \propto \sigma_b^{-(n+\nu_b)}\exp\Big[-\frac{1}{2}\sigma_b^{-2}(\sum_{i=1}^{n}{b_i}^2+ss_b)\Big] \tag{55}$$

and hence, we can sample $\sigma_b^2$ from an inverse gamma distribution with scale parameter $(\sum {b_i}^2+ss_b)/2$ and degrees of freedom $(n+\nu_b)/2$.

**Sampling of $\boldsymbol{\beta}$**
The distribution from which we can sample $\boldsymbol{\beta}$ is slightly different than in a normal linear regression model, due to the random effect specification. The full conditional posterior of $\boldsymbol{\beta}$ is given by:

$$p(\boldsymbol{\beta}|\cdot) \propto \exp\Big[-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-b_i)'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-b_i)\Big]\exp\Big[-\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\Big], \tag{56}$$

with $\boldsymbol{V}_i = \sigma_b^2\boldsymbol{J}_{T_i}+\sigma_\varepsilon^2\boldsymbol{I}_{T_i}$. From the conditional distribution it follows that we can sample $\boldsymbol{\beta}$ from a normal distribution with mean $\sum_{i=1}^{n}(\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i+\boldsymbol{S}_0^{-1})^{-1}(\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{y}_i+\boldsymbol{S}_0^{-1}\boldsymbol{\beta}_0)$ and variance $(\boldsymbol{V}_i+\boldsymbol{S}_0^{-1})^{-1}$.

**Sampling of $\boldsymbol{\upsilon}, \vartheta$ and $\sigma_\eta^2$**
The sampling of the parameters corresponding to initial response equation, $\boldsymbol{\upsilon}$, $\vartheta$ and $\sigma_\eta^2$, is relatively straightforward, as these parameters can be treated as parameters from an ordinal linear regression model. Using that the regression equation is given by

$$y_{i0} = \boldsymbol{z}_i'\boldsymbol{\upsilon}+b_i\vartheta+\eta_i, \tag{57}$$

we can use that our predictor variables are given by $\dot{\boldsymbol{z}}_i = (\boldsymbol{z}_i', b_i)'$ and our depedent variable is given by $y_{i0}$. Conditional on the random effects in $\boldsymbol{b}$, the regression parameters in $\dot{\boldsymbol{\upsilon}} = (\boldsymbol{\upsilon}', \vartheta)'$ and the residual variance parameter can be sampled from well-known posteriors of an ordinary linear regression model with conjugate priors, which can be found in various introductory textbooks on Bayesian statistics, such as Greenberg (2012). After stacking the initial responses and corresponding external covariates of all $n$ individuals, we can sample the parameter vector $\dot{\boldsymbol{\upsilon}}$ from a normal distribution with mean $(\sigma_\eta^{-2}\dot{\boldsymbol{Z}}'\dot{\boldsymbol{Z}}+\boldsymbol{U}_0^{-1})^{-1}(\sigma_\eta^{-2}\dot{\boldsymbol{Z}}'\boldsymbol{y}_0+\boldsymbol{U}_0^{-1}\boldsymbol{\upsilon}_0)$ and variance $(\sigma_\eta^{-2}\dot{\boldsymbol{Z}}'\dot{\boldsymbol{Z}}+\boldsymbol{U}_0^{-1})^{-1}$.

Similarly, the residual variance of the initial response equation $\sigma_\eta^2$ can be treated as the residual variance from an ordinary linear regression model. Hence, we can sample $\sigma_\eta^2$ from an inverse gamma distribution with scale parameter $((\boldsymbol{y}_0 - \dot{\boldsymbol{Z}}'\dot{\boldsymbol{\upsilon}})'(\boldsymbol{y}_0 - \dot{\boldsymbol{Z}}'\dot{\boldsymbol{\upsilon}}) + ss_\eta)/2$ and degrees of freedom equal to $(n + \nu_\eta)/2$.

**Sampling of $\boldsymbol{b}$**

When we rearrange the terms from the initial observation, as given in Equation (8), and divide everything by the standard deviation of the error term, we obtain:

$$(y_{i0} - \boldsymbol{z}_i'\boldsymbol{\upsilon})/\sigma_\eta = b_i\vartheta/\sigma_\eta + \eta_i/\sigma_\eta \tag{58}$$

Similairly, subsequent observations of individual $i$ can be rewritten as:

$$(y_{it} - \boldsymbol{x}_{it}'\boldsymbol{\beta})/\sigma_\varepsilon = b_i 1/\sigma_\varepsilon + \varepsilon_{it}/\sigma_\varepsilon \tag{59}$$

Then, in order to derive the sampling distribution of the random effects we can stack the observations $0, ..., T_i$ for every individual, resulting in the two auxilliary variables:

$$\boldsymbol{y}_i^* = \begin{pmatrix} (y_{i0} - \boldsymbol{z}_i'\boldsymbol{\upsilon})/\sigma_\eta \\ (y_{i1} - \boldsymbol{x}_{i1}'\boldsymbol{\beta})/\sigma_\varepsilon \\ \vdots \\ (y_{iT_i} - \boldsymbol{x}_{iT_i}'\boldsymbol{\beta})/\sigma_\varepsilon \end{pmatrix}, \quad \boldsymbol{x}_i^* = \begin{pmatrix} \vartheta/\sigma_\eta \\ 1/\sigma_\varepsilon \\ \vdots \\ 1/\sigma_\varepsilon \end{pmatrix} \tag{60}$$

After defining these two auxiliary variables, we can use the results from an ordinary linear regression model with standard normally distributed error term. We treat the random effects $b_i$ as our regression parameter, $\boldsymbol{y}_i^*$ as our dependent variable and $\boldsymbol{x}_i^*$ as our predictor. Due to the conjugacy of the prior specification, we know that we can sample each random effect $b_i$ from a normal distribution with mean $(\boldsymbol{x_i^{*\prime}}\boldsymbol{x_i^*} + \sigma_b^{-2})^{-1}\boldsymbol{x_i^{*\prime}}\boldsymbol{y_i^*}$ and variance $(\boldsymbol{x_i^{*\prime}}\boldsymbol{x_i^*} + \sigma_b^{-2})^{-1}$.

# C Supplementary results of the multivariate analysis

**Table 16:** Posterior results of covariance matrix and correlation of the residuals of the regression analysis of Equation (19)

| | **Females** | | | **Males** | | |
|---|---|---|---|---|---|---|
| | mean | 95% CI | | mean | 95% CI | |
| $\sigma^2_{\varepsilon 1}$ | 0.186 | 0.176 | 0.195 | 0.218 | 0.207 | 0.230 |
| $\sigma^2_{\varepsilon 2}$ | 0.028 | 0.026 | 0.029 | 0.028 | 0.027 | 0.030 |
| $\sigma_{\varepsilon 21}$ | 0.003 | 0.000 | 0.005 | 0.004 | 0.001 | 0.007 |
| $\rho(\varepsilon_1, \varepsilon_2)$ | 0.038 | 0.003 | 0.072 | 0.054 | 0.018 | 0.089 |

**Table 17:** Posterior results of multivariate regression of Equation (19) for males.

| | Hb | | | ZPP | | |
|---|---|---|---|---|---|---|
| | mean | 95% CI | | mean | 95% CI | |
| (Intercept) | 7.414* | 6.812 | 8.029 | 2.814* | 2.572 | 3.059 |
| Age | −0.006* | −0.008 | −0.003 | −0.003* | −0.004 | −0.001 |
| Spring | −0.020 | −0.064 | 0.024 | −0.062 | −0.078 | −0.002 |
| Summer | 0.020 | −0.024 | 0.064 | −0.079* | −0.096 | −0.064 |
| Autumn | −0.010 | −0.051 | 0.034 | 0.040 | 0.024 | 0.057 |
| Time of the day | −0.007* | −0.012 | −0.002 | 0.000 | −0.002 | 0.002 |
| BMI | 0.013* | 0.001 | 0.025 | 0.013* | 0.007 | 0.019 |
| Bloodvolume | 0.047 | −0.025 | 0.120 | −0.026 | −0.061 | 0.009 |
| Number of previous donations last 2 years | 0.008 | −0.002 | 0.018 | 0.021* | 0.018 | 0.025 |
| Time since previous donation (months) | 0.005* | 0.001 | 0.011 | −0.010* | −0.012 | −0.008 |
| Previous Hb value | 0.130* | 0.095 | 0.165 | 0.013 | −0.002 | 0.024 |
| Previous ZPP value | 0.105 | −0.016 | 0.204 | 0.241* | 0.201 | 0.281 |

The asterisk (*) denotes that zero is not included in the 95% highest posterior density interval.

**Table 18:** Posterior results of multivariate regression of Equation (19) for females.

| | Hb | | | ZPP | | |
|---|---|---|---|---|---|---|
| | mean | 95% CI | | mean | 95% CI | |
| (Intercept) | 7.132* | 6.641 | 7.634 | 3.362* | 3.118 | 3.557 |
| Age | 0.003 | 0.000 | 0.006 | −0.085 | −0.100 | −0.070 |
| Spring | −0.012 | −0.051 | 0.027 | −0.135 | −0.150 | −0.120 |
| Summer | −0.024 | −0.062 | 0.015 | 0.094* | 0.087 | 0.102 |
| Autumn | 0.001 | −0.039 | 0.041 | 0.020 | 0.005 | 0.035 |
| Time of the day | −0.001 | −0.005 | 0.004 | 0.001 | −0.001 | 0.002 |
| BMI | 0.003 | −0.006 | 0.012 | 0.007* | 0.002 | 0.011 |
| Bloodvolume | 0.047 | −0.026 | 0.120 | 0.004 | −0.038 | 0.045 |
| Number of previous donations last 2 years | 0.000 | −0.014 | 0.015 | 0.024* | 0.018 | 0.029 |
| Post menopause | 0.093* | 0.017 | 0.169 | 0.004 | −0.032 | 0.039 |
| Time since previous donation (months) | 0.000 | −0.005 | 0.004 | −0.013* | −0.014 | −0.011 |
| Previous Hb | 0.094* | 0.063 | 0.125 | 0.002 | −0.009 | 0.013 |
| Previous ZPP | −0.049 | −0.040 | 0.137 | 0.210* | 0.172 | 0.248 |

The asterisk (*) denotes that zero is not included in the 95% highest posterior density interval.

# D Full list of available variables

**Table 19:** Variables used in primary models

| Variable name | Description |
|---|---|
| Donor ID | Number that defines unique individual |
| Number of previous donations | Number of previous donations in the last 2 years |
| Age | Age at time of the donation |
| Date | Date at which the donation took place |
| Time | Time of the day at which the donation took place |
| ZPP | Zinc Protoporphyrin (µmol/mol heme) |
| Weight | Weight of donor in kg (only measured at first visit) |
| Height | Height of donor in cm (only measured at first visit) |
| HGB | hemoglobin concentration (mmol/l) |

**Table 20:** Additional explanatory variables describing blood levels

| Variable name | Description |
|---|---|
| WBC | white bloodcell count ($10^9$/l) |
| RBC | red bloodcell count ($10^{12}$/l) |
| HCT | hematocrit (%) |
| MCV | mean corpuscular volume (fl) |
| MCH | mean corpuscular haem (fmol) |
| MCHC | mean corpuscular haem concentration (mmol/l) |
| PLT | platelet count ($10^9$/l) |
| RDWSD | red cell diameter width standard deviation (fl) |
| RDWCV | red cell diameter width coefficient of variation (%) |
| PDW | platelet distribution width (%) |
| MPV | mean platelet volume (fl) |
| PLCR | platelet-large cell ratio (%) |
| PCT | plateletcrit (%) |
| NEUT | neutrophil count ($10^9$/l) |
| LYMPH | lymphocyte count ($10^9$/l) |
| MONO | monocyt count ($10^9$/l) |
| EO | eosinophil count ($10^9$/l) |
| BASO | Basophil count ($10^9$/l) |
| IG | immature granulocyte count ($10^9$/l) |
| Systolic blood pressure | Pressure in blood vessels during heartbeat (mmHg) |
| Diastolic blood pressure | Pressure in blood vessels during rest (mmHg) |