Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Oliver Byhring

Relative variable importance: A comparison between R^2 decomposition and variable importance in machine learning

Master's thesis in Applied Physics and Mathematics
Supervisor: Stefanie Muff
March 2021

NTNU
Norwegian University of Science and Technology

Oliver Byhring

Relative variable importance: A comparison between R^2 decomposition and variable importance in machine learning

Master's thesis in Applied Physics and Mathematics
Supervisor: Stefanie Muff
March 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

When doing regression analysis, we are often interested in what predictors have the strong influence on the response. While a lot of research has been done in this context on linear regression models, there is still a lot to explore in mixed-effect models. It is common in linear regression models that the importance of the predictors should be a decomposition of the variance explained by the model. In mixed-effect models it is not immediately clear what proportion of variance is explained by the fixed-effects and what is explained by the random-effects.

This thesis aims to discuss the extension of an existing method of assigning relative importance in linear regression models and compare the new extended method to variable importances assigned by a random forest method. The methods will be illustrated on two examples, namely a simulated data set and a study of children's activity level (SPLASHY).

Random forests are a statistical learning method that naturally can provide a relative importance measure. Although random effects in trees are not so straightforward, it is possible to encode a random effect variables as a categorical variables to make the trees handle the random effects. The variable importance estimate from the random forests can then be used as a comparison for the relative variable importance metric in random intercept models. However, since the importances assigned in a random forest does not decompose a model statistic and the magnitude of the importances depend on the scale of the response, the importances are standardized before comparison.

The existing method for assigning relative variable importance in a regular linear regression models, called the LMG-method, requires a goodness-of-fit measure. It is common to use the explained variance, R^2. However, for linear mixed models, there are several ways to define R^2. Most importantly, a distinction is made between marginal and conditional R^2 where the marginal considers only the variance explained by the fixed predictors and the conditional considers the variance explained by the random intercept in addition to the fixed predictors.

An `R` package with functions to calculate the relative importance in random intercept models is also a product of this thesis. How to install and use it is described in Appendix A.

Sammendrag

I en regresjonsanalyse er vi ofte interessert i hvilke parametere som har størst påvirkning på responsvariabelen. Selv om det er gjort mye forskning på dette området når det kommer til lineære regresjonsmodeller, er det fortsatt en del som kan utforskes når det kommer til blandede modeller. I lineære regresjonsmodeller er det vanlig at viktigheten til parameterene er en dekomposisjon av variansen som er forklart av modellen. I blandede modeller er det ikke åpenbart hvor stor andel av variansen som er forklart av faste effekter og hvor stor del som er forklart av blandede effekter.

Formålet med denne avhandlingen er å diskutere en utvidelse av en eksisterende metode for å bestemme relative viktighet i lineære regresjonsmodeller, og sammenligne den utvidede metoden med relativ viktighet fra random forests. Metoden vil bli illustrert på to eksempler, et simulert datasett og en studie av aktivitetsnivået til barn (SPLASHY).

Random forests er en statistisk læringsmetode som naturlig kan gi et mål på relativ viktighet. Selv om det å håndtere blandede effekter i trær ikke er helt rett frem, er det mulig å kode blandede effekt variable som kategoriske variable for å gjøre de mer håndterbare for trær. Estimatet av variabelviktighet fra random forests kan da bli brukt som en sammenligning for det relative variabelviktighetsmålet fra de blandede modellene. Siden viktighetene som blir tildelt variablene i en random forest ikke dekomponerer en modellstatistikk, og størrelsen på viktighetene avhenger av skalaen til responsen, blir viktighetene standardiserte før sammenligning.

Den eksisterende metoden for å tildele relativ variabelviktighet i vanlige lineære modeller, kalt LMG-metoden, krever et godhetsmål (goodness-of-fit) på modellen. Det er vanlig å bruke forklart varians, R^2. For blandede lineære modeller er det i midlertid flere måter man kan definere R^2. Viktigst er skille mellom marginal og betinget R^2, hvor marginal kun tar hensyn til variansen forklart av de faste effektene, mens betinget tar hensyn til variansen forklart av både de blandede og de faste effektene.

En `R` pakke med funksjoner for å beregne de relative viktighetene i tilfeldig skjæringspunktmodeller er også et produkt av denne avhandlingen. En beskrivelse for hvordan installere og bruke denne pakken finnes i Appendix A.

Preface

This master's thesis was written during the last semester of my Master of Technology degree at the Norwegian University of Science and Technology (NTNU). This final assignment marks the end of the study programme *Applied Physics and Mathematics*, with specialization in *Industrial Mathematics*.

I want to give special thanks to my supervisor Stefanie Muff for excellent guidance.

Oliver Byhring
Trondheim, 12.03.2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Linear regression models are frequently used in statistical analyses. One of these models' main objective is to explain the influence of a set of regressors on a response variable. In this context, it can be interesting to know how influential each predictor is relative to each other. There are several ways relative importance can be defined, and for linear models an intuitive metric is to take the size of the (standardized) coefficients or the p-value of the coefficient. However, these metrics have some limitations for assessing relative variable importance.

In recent years tree-based models have seen an increase in popularity due to their ability to map non-linear relationships with high accuracy. Random forest is a tree-based method that can handle both regression and classification problems using either continuous or categorical regressor variables. One positive aspect of random forests models is that they can provide an estimate of the variable importance of the variables in the model (Breiman, 2001; Strobl et al., 2008; Zhu et al., 2015).

The topic of how to assign variable importance has been widely discussed in linear regression literature (e.g., Pratt, 1987; Kruskal and Majors, 1989; Liu et al., 2014; Grömping, 2015). A straightforward idea is to decompose the total variance explained by the model, R^2, into shares explained by the individual predictors. Several studies has also been done on relative importance in random forests (Breiman, 2001; Strobl et al., 2008; Zhu et al., 2015; Gregorutti et al., 2017). In random forests, on the other hand, it is common to assign importance to the predictors equal to the predictor's contribution to the reduction in error (James et al., 2013). The magnitude of the importances, therefore, depends on the scale of the response variable.

Decomposing the variance of a linear model is simple when the covariates in the model are uncorrelated, but it is not obvious when the covariates are correlated with each other, which they typically are in most real-world applications. This is because the covariates absorb variance from each other in particular when one is not included in the model. Lindeman et al. (1980) suggested a method for assigning relative variable importance in regular linear regression models when regressors are correlated. The method is often referred to as the LMG-method due to the initials of the authors. This method is reviewed in detail by Grömping (2015), which also compares the method

with other common methods of assigning variable importance. It is the LMG-method that will be the main focus of this thesis.

The LMG-method has previously been used to assign importances to the regressors in linear regression models by decomposing the explained variance of the full regression model (Grömping, 2007). The relative importance assigned to a predictor represents the amount of variance explained by that predictor. The way Lindeman et al. (1980) assign relative importance is by looking at all possible permutations of the regressors in all the subset models of the full model, where the assigned importance to a predictor is the average incremental goodness of fit when a regressor is added to a subset-model. This way of assessing relative importance is by several researchers regarded as a comprehensive approach for determining variable importance (Grömping, 2015; Cançado, 2018). The largest relative variable importance is assigned to the variable that gives, on average, the biggest increment to the assessment of the model fit. The LMG-method is, however, restricted to regular linear regression models.

It is not uncommon in applications that observations are correlated. In medical and biological analysis there is often more than one measurement per subject, this is referred to as repeated measures. Mixed-effects models are designed to handle these types of data by letting observations of one subject get their own intercept and/or slope. If a regular linear model is fitted, the model residuals would be treated as independent, which they are not if two or more observations are correlated, resulting in too small standard errors in the model coefficients.

While the main focus of earlier papers has been on assigning variable importances in regular linear models, the topic of assigning relative importance in mixed-effect models has lacked attention. As of today, there is no universal agreement on how to assign relative importances in mixed-effect models, and there is still some work to be done. Liu et al. (2014) have explored the use of the Pratt index, a method suggested by Pratt (1987). The Pratt index is a share assigned to a regressor equal to the standardized regression coefficient corresponding to the regressor times the correlation between the regressor and the response. Grömping (2015) mentions the Pratt index as too simple and argues that the Pratt index violates essential properties that a variable importance metric should possess, such as avoiding to assign negative importances. Byhring (2020) proposed a way of extending the LMG-method to handle random intercept models, this method is described in Chapter 2. It is of interest that the relative importance metric of a regressor should, as in the regular linear regression models, reflect the proportion of variance explained by the respective regressor.

While it is common in a regular linear regression model to do an R^2 decomposition into shares that are assigned to each of the variables, it is not directly obvious how to define R^2 in mixed models. It is in principle possible to do a decomposition of the Akaike Information Criterion (AIC) or the Bayesian information criterion (BIC), as these methods are used to assess the model-fit of mixed effect models. However, AIC and BIC only provide an estimate of the relative fit in comparison with other model-fits. The relative importance assigned to a regressor would then not have a practical meaning in comparison to the R^2 decomposition in regular linear regression models. It would therefore be desirable to have a measure of the explained variance

in a mixed effect model. Nakagawa and Schielzeth (2013) have proposed several ways of extending the explained variance in regular linear regression models to linear mixed models and generalized linear mixed models. This will be discussed further in Chapter 2, where we outline the theory.

Chapter 2 presents the relevant theory and begins by introducing the linear regression model and a common definition of R^2 for this model. We then go on to mention some useful properties a variable importance metric should possess and defines the LMG-metric for the linear regression model before we introduce the random intercept model and look at how R^2 can be defined for this type of models. Several possible definitions of relative importance in random intercept models are provided in Chapter 2. We present some tree-based methods and describe how these methods handle categorical variables, along with relative importance metrics for these methods. The theory is concluded with a section on how to sample a synthetic data set where it is easy to control the parameters of a random intercept model. In Chapter 3 we look at two data sets, fit different statistical models to the data sets and calculate the relative importances. Chapter 4 contains a comparison of the assigned importances from the random intercept models and random forest models, and we discuss the performance of the suggested extension to the LMG-method. The R code used to implement the methods will be provided in the Appendix.

Chapter 2

Theory

2.1 Linear regression model and R^2

When there is a linear relationship between the predictors, $\boldsymbol{X}$, and the response, $\boldsymbol{Y}$, it is possible to use a linear model to do regression analysis. Linear regression is a form of supervised learning. That is, the response is known and used to improve the model fit. This is a simple yet powerful way to predict the quantitative relationship between the response and the predictors. We begin by looking at a standard linear regression model of the form

$$y_i = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i + \varepsilon \ , \tag{2.1}$$

where y_i is the i^{th} response, β_0 is the model intercept, $\boldsymbol{\beta}^T$ is the vector of fixed slopes corresponding to the covariates $\boldsymbol{x}_i$ with elements $(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(p)})$ and ε is the error term, which is a collection of the information the model is unable to catch due to its simplicity. The error term is assumed to be normally distributed with mean zero, variance σ_ε^2 and is independent of $\boldsymbol{x}$. Since there are p covariates, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector and assuming y is a $(1 \times n)$ vector, then $\boldsymbol{x}_i$ is a $(p \times n)$ matrix.

The model used to make predictions can be defined as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i \ ,$$

where $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ are the estimated coefficients of the model. It is common to refer to difference between the observed and the predicted values as residuals, $e = (e_1, \cdots, e_n)$. The i^{th} residual is defined as

$$e_i = y_i - \hat{y}_i \ .$$

The residuals, e_i, are assumed to be independent and identically distributed

$$e_i \sim \mathcal{N}(0, \sigma_\varepsilon^2) \ .$$

The coefficients $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}^T$ are chosen such that the sum of the squared residuals,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ , \tag{2.2}$$

is minimized.

The coefficient of determination, R^2, is a measure of the proportion of variance explained by all the covariates in the model. This measure is commonly used to evaluate the goodness of fit in linear regression models. The R^2 takes values between 0 and 1, and the closer the value is to 1, the more of the variance is explained by the model. The variance of the response Y can be decomposed as

$$\text{Var}(Y) = \text{Var}(\hat{Y}) + \text{Var}(e),$$

since the residuals, e, are assumed to be independent of the covariates, $\boldsymbol{x}$. The proportion of variance that is explained can be expressed as

$$R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = \frac{\text{Var}(Y) - \text{Var}(e)}{\text{Var}(Y)} = 1 - \frac{\text{Var}(e)}{\text{Var}(Y)} \ . \tag{2.3}$$

This expression for the variance explained can be interpreted as 1 minus the variance unexplained.

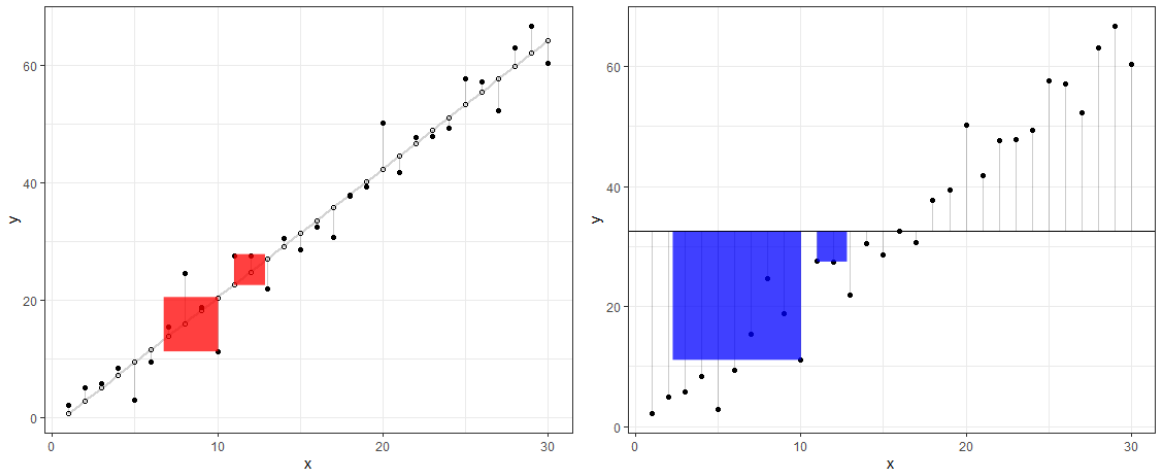

Figure 2.1: The left plot shows a linear model fitted to the data and the red squares visualizes the residuals $(y_i - \hat{y}_i)^2$ for the tenth and eleventh data point. The right plot shows the mean of the data, $\overline{y}$, as a horizontal line and the blue squares shows the residuals $(y_i - \overline{y})^2$ for the tenth and eleventh data point.

Figure 2.1 illustrates the squared residuals $(y_i - \hat{y}_i)^2$ and the total squares $(y_i - \overline{y})^2$. The sum of squared residuals is often referred to as RSS, and the total sum of squares is referred to as TSS. It is possible to rewrite equation (2.3) in terms of the RSS and TSS. This leads to the most common definition of R^2 for a standard linear model,

$$R^2_{\text{LM}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \ ,$$

where $\overline{y}$ is the grand mean of the response and the subscript LM signifies that the model is a regular linear model as defined in equation (2.1). For the R^2 to take values close to 1 it is necessary for the sum of squared residuals, defined in equation (2.2), to be as small as possible. If the value of R^2 is exactly one, the model fits the data perfectly and all residuals are $e_i = 0$.

2.2 Relative importance in linear regression models

It is sometimes of interest to find out what proportion of the variance of the response, Y, is explained by each of the predictors, X's, instead of the ensemble of all predictors. The proportion of variance explained will represent its relative importance.

The expression for the variance of the response, Y, can be written

$$\begin{aligned} \text{Var}(Y) &= \text{Var}\left(\sum_{k=1}^{p} \beta_k x^{(k)}\right) + \sigma_\varepsilon^2 \\ &= \sum_{i=1}^{p} \beta_i^2 \alpha_i + 2 \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} \beta_k \beta_l \rho_{kl} + \sigma_\varepsilon^2 \ , \end{aligned} \tag{2.4}$$

where α_i denotes the variance of $x^{(i)}$ and $\rho_{i,j}$ denotes the covariance between x_i and x_j. From equation (2.4) we see that when the regressors are uncorrelated the variance simply becomes

$$\text{Var}(Y) = \sum_{i=1}^{p} \beta_i^2 \alpha_i + \sigma_\varepsilon^2 \ .$$

In this situation decomposing the variance of the response becomes very easy. It simply decomposes into the contributions $\beta_i^2 \alpha_i$. On the other hand, when the regressors are correlated, it is not obvious how the different regressors contribute to R^2. The reason for this is that when regressors are correlated they absorb variance from each other when one is not included in the model.

When defining a variable importance metric it is often of interest that the metric has certain properties. Grömping (2015) lists several properties for a relative variable importance metric which are considered useful in the literature. The five most relevant metrics are the following:

- *Proper decomposition*: the model variance is decomposed, that is, the sum of all shares is the model variance (or R^2, depending on normalization).

- *Orthogonal compatibility*: The decomposition respects orthogonal subgroups, i.e., for each orthogonal subgroup of regressors, the assigned shares sum to the unique overall model variance (or R^2) for that subgroup.
- *Non-negativity*: all allocated shares are always non-negative.
- *Exclusion*: the share allocated to a regressor X_j with $\beta_j = 0$ should be 0.
- *Inclusion*: a regressor X_j with $\beta_j \neq 0$ should receive a non-zero share.

It is a common requirement that the metric should be properly decomposed into non-negative shares, as this is a common request from customers of statistical analysis (Grömping, 2007). Hence these two properties (proper decomposition and non-negativity) will be of main interest when extending the LMG-method to random intercept models later in this thesis. Exclusion is often regarded as an undesirable criterion when the relative importance question is asked with a causal interpretation (Grömping, 2007). There are methods for assigning relative importance in linear regression models that are designed to fulfill the exclusion criterion, such as the proportional marginal variance decomposition (PMVD) (Feldman, 2005). The downside to the methods designed to fulfill the exclusion criterion is that they often require increased computation effort and are harder to implement.

The LMG-method, as suggested by Lindeman et al. (1980), revolves around permuting variables in subset models of the full model and then looking at the increment in R^2 when a regressor is added to the model. It is, therefore, useful to introduce some notation that will simplify calculations. The regressors will be labeled and denoted $X^{(1)}, \cdots, X^{(p)}$. The order of which regressors are entered into the model is denoted $r = (r_1, \cdots, r_p)$, which is a permutation of the regressors with indices $\{1, \cdots, p\}$. The set of regressors that appears before $X^{(1)}$ in permutation r is denoted $S_1(r)$. In general, we have that the set of regressors that appear before the i^{th} regressor, $X^{(i)}$ in permutation r is denoted $S_i(r)$.

Grömping (2007) defines evar(.) and svar(.) to further simplify notation for the calculations

$$\begin{aligned} \text{evar}(S) &= \text{Var}(Y) - \text{Var}(Y|X_j, j \in S) \\ \text{svar}(M|S) &= \text{evar}(M \cup S) - \text{evar}(S) \ , \end{aligned}$$

where evar(.) denotes the explained variance of a model with regressors from the set S of regressors and svar(.) denotes the increase in explained variance when adding the regressors from the set M of regressors to a model that already contains the regressors from the set S.

The importance assigned to a regressor is equal to the average increment in R^2 over all possible permutations of regressors, when adding the regressor to the model. Without loss of generality Grömping (2007) defines the LMG for the first predictor, $X^{(1)}$, as

$$\text{LMG}(1) = \frac{1}{p!} \sum_{\pi \text{ permutations}} \text{svar}(\{1\}|S_1(\pi)),$$

but this can easily be generalized to the i^{th} predictor, $X^{(i)}$, as

$$\text{LMG}(i) = \frac{1}{p!} \sum_{\pi \text{ permutations}} \text{svar}(\{i\}|S_i(\pi)), \tag{2.5}$$

This is an unweighted sum of all orderings that contribute to the relative importance metric for regressor i. To get a more intuitive understanding of the expression for LMG in equation (2.5), Berg (2019) has written the expression in terms of R^2,

$$\begin{aligned}
\text{LMG}(i) &= \frac{1}{p!} \sum_{\pi \text{ permutations}} \text{svar}(\{i\}|S_i(\pi)) && (2.6)\\
&= \frac{1}{p!} \sum_{\pi \text{ permutations}} (\text{evar}(\{i\} \cup S_i(\pi)) - \text{evar}(S_i(\pi))) && (2.7)\\
&= \frac{1}{p!} \sum_{\pi \text{ permutations}} \Big(R^2\big(\{i\} \cup S_i(\pi)\big) - R^2\big(S_i(\pi)\big) \Big). && (2.8)
\end{aligned}$$

In equation (2.8), the notation R^2 of a set of indices means R^2 of the regular linear regression model with the regressors corresponding to the indices.

We see that this is simply the average increase in the model R^2 when adding $X^{(i)}$ to the model in the respective permutation order averaged over all possible permutations. Note that there are $p!$ possible permutations, thus $\pi = \{1, \cdots, p!\}$.

The order of the regressors that appear before a regressor, $X^{(i)}$, a permutation, π, is irrelevant for the model fit, the same holds for the regressor that appears after $X^{(i)}$. To illustrate this, consider the equivalent models

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i \ , \\
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_5 x_{i5} + \beta_4 x_{i4} + \varepsilon_i \ , \\
y_i &= \beta_0 + \beta_2 x_{i2} + \beta_1 x_{i1} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i \text{ and} \\
y_i &= \beta_0 + \beta_2 x_{i2} + \beta_1 x_{i1} + \beta_3 x_{i3} + \beta_5 x_{i5} + \beta_4 x_{i4} + \varepsilon_i \ .
\end{aligned} \tag{2.9}$$

Regardless of the order of the regressors that appear before and after $X^{(3)}$, the four linear models in equation (2.9) are the same. The models, therefore, also have equal R^2. This can be used to reduce the number of required computations in equation (2.8). Equation (2.8) can be written

$$\text{LMG}(i) = \frac{1}{p!} \sum_{S \subseteq (2, \cdots, p)} n(S)!(p - n(S) - 1)! \Big(R^2\big(\{i\} \cup S\big) - R^2\big(S\big) \Big) \ , \tag{2.10}$$

where $n(S)!$ is the number of possible permutations of the predictors that appears before $X^{(i)}$ and $(p-n(S)-1)!$ is the number of possible permutations of the predictors that appear after $X^{(i)}$. This reduces the numbers of summands required to compute from $\pi!$ to $2^{\pi-1}$.

2.3 Random intercept models and R^2

To model clustered or grouped data, for example in the presence of repeated observations on the same individual, we introduce a random intercept model. In the case of random intercept models, the response, Y will have two indices, one specifying the individual number, i, and one for the observation number, j. The random intercept accounts for the fact that observations within the same individual i, y_{ij} and y_{ik}, are correlated. The random intercept ensures that between two individuals i and r the observations y_{ij} and y_{rj} are uncorrelated. The random intercept, γ_i, can be interpreted as i^{th} individual's deviation from the population mean.

The model equation for the j-th observation of the i^{th} individual is

$$y_{ij} = \beta_0 + \gamma_i + \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \varepsilon \ , \tag{2.11}$$

where β_0 is the fixed population intercept, $\boldsymbol{\beta}$ is the the $(1 \times p)$ vector of fixed population slopes of the covariates $\boldsymbol{x}_{ij}$ and γ_i is the individual specific deviance from the population intercept. The error, ε, and the individual specific deviances from the population intercept is assumed to be independent from $\boldsymbol{x}$ and each other and normally distributed with mean zero and variance σ_ε^2 and σ_γ^2 respectively, that is,

$$\begin{aligned} \varepsilon &\sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right) \\ \gamma_i &\sim \mathcal{N}\left(0, \sigma_\gamma^2\right) \ . \end{aligned}$$

Let ρ_{ij} be the covariance between the fixed regressors $X^{(i)}$ and $X^{(j)}$. Then for the random intercept model defined in equation (2.11), the variance from the fixed effect can be expressed as

$$\begin{aligned} \sigma_f^2 &= \operatorname{Var}\left(\sum_{k=1}^{p} \beta_k x^{(k)}\right) \\ &= \sum_{i=1}^{p} \beta_i^2 \alpha_i + 2 \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} \beta_k \beta_l \rho_{kl}, \end{aligned} \tag{2.12}$$

where α_i is the variance of the i^{th} covariate. This is equivalent to the expression for the variance of the response in regular linear regression models, as described in equation (2.4), just without the variance from the residuals, σ_ε^2. Thanks to the independence assumptions for γ_i and ε_{ij} the variance of Y can be written as a sum of all variance components, that is,

$$\operatorname{Var}(Y) = \sigma_f^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2 \ .$$

Defining R^2 in a random intercept model is not as straightforward as one might first think. It is therefore not uncommon that information criteria are used as comparison

tools for mixed models. Information criteria are methods that evaluate the probability of the data given the fitted model. However, there are several limitations to using information criteria. They do not give any information about the overall goodness of model fit, and they also provide no information about how much of the variance is explained by the model (Nakagawa and Schielzeth, 2013). It is therefore of interest to find a way to generalize R^2 to random intercept models.

When defining R^2 in random intercept models, a choice has to be made whether to define R^2 as the variance explained by the fixed effects alone, or the variance explained by the random and fixed effects combined. Nakagawa and Schielzeth (2013) have proposed several ways of defining R^2 in linear mixed models and generalized linear mixed models. In particular, Nakagawa and Schielzeth (2013) distinguish between *marginal* R^2, denoted $R^2_{\text{LMM(m)}}$, and *conditional* R^2, denoted $R^2_{\text{LMM(c)}}$.

$R^2_{\text{LMM(m)}}$ is the proportion of variance explained by the fixed effect components alone. The expression for the marginal R^2 is

$$R^2_{\text{LMM(m)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \ . \tag{2.13}$$

An advantage of this definition of R^2 in random intercept models is that it never becomes negative, in contrast to other proposed definitions of R^2 (Nakagawa and Schielzeth, 2013). It can occur that R^2 decreases when a new variable is added, although, this is unlikely as the variance explained by the fixed effects, σ_f^2, always increases when a new variable is added.

Nakagawa and Schielzeth (2013) also defined the conditional R^2 as the variance explained by both the fixed effects and the variance of the random effects. In a random intercept model, the only variance from the random effects is the random intercept variance, σ_γ^2. Thus the expression for the conditional R^2 is

$$R^2_{\text{LMM(c)}} = \frac{\sigma_f^2 + \sigma_\gamma^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \ .$$

2.4 Relative importance in random intercept regression model

There are different approaches to assigning importance to the regressors when we generalize the LMG approach from linear models to linear mixed models. The two main questions are:

- Are we going to assign importance to the random intercept?
- Should the random intercept be permuted as the other regressors?

If the random intercept is to be assigned an importance like the other variables, then a possible approach is to permute it in the same way as the regressors. The LMG

expression for the i^{th} regressor, $X^{(i)}$, then becomes

$$\text{LMG}(i) = \frac{1}{(p+1)!} \sum_{S \subseteq (1,\cdots,p,RI)\setminus i} n(S)!(p-n(S))! \Big(R^2\big(\{i\} \cup S\big) - R^2\big(S\big) \Big) \ , \quad (2.14)$$

where RI stands for random intercept variable. The difference between this expression and the LMG-expression for regular linear regression, described in equation (2.10) is that there is now $p+1$ variables that are to be permuted. Notice that the random intercept is not always in the set S. That means the contributions sometimes are the increase in R^2 in a regular linear model when adding the predictor to the model, other times it is the increase in R^2 in a random intercept model.

From equation (2.14) it is evident that the LMG expression for the random intercept becomes

$$\text{LMG}(RI) = \frac{1}{(p+1)!} \sum_{S \subseteq (1,\cdots,p)} n(S)!(p-n(S))! \Big(R^2\big(\{RI\} \cup S\big) - R^2\big(S\big) \Big) \ . \quad (2.15)$$

Observe that in equation (2.15) the share assigned to the random intercept is the average increase in R^2 when adding the random intercept to the models, thus always comparing a random intercept model with a regular linear model. This might cause the proper decomposition criteria to be violated. Another possible solution to overcome the issue of improper decomposition is to not permute the random intercept but rather let it be in the model in all permutations of the covariates. The LMG expression for the i^{th} regressor, $X^{(i)}$, then becomes

$$\text{LMG}(i) = \frac{1}{p!} \sum_{S \subseteq (1,\cdots,p)\setminus i} n(S)!(p-n(S)-1)! \Big(R^2\big(\{i\} \cup S\big) - R^2\big(S\big) \Big) \ .$$

In this situation, the random intercept does not get assigned an importance, but the importances should decompose the explained variance properly with no negative shares regardless of whether R^2_c or R^2_m is used to compute the explained variance. If R^2_c is used, then the shares assigned to regressors are expected to be artificially high when not assigning any importance to the random intercept. The reason is that the random intercept variance is then (wrongly) interpreted as variance explained by the fixed effects because the random intercept is already in the model when the first predictor is added. Using R^2_m defined in equation (2.13), instead of R^2_c when calculating the importances is expected to result in more realistic shares being assigned to the regressors. The proper decomposition holds in both scenarios. However, when R^2_m is used to assess the model-fit, the relative importances sum up to the variance explained by the fixed effects alone.

It is meaningful to assign an importance to the random intercept equal to the difference of the marginal- and conditional R^2 of the full model since R^2_m will always be smaller

than R_c^2. The random intercept importance can then be defined as

$$\text{LMG}(RI) = R_c^2 - R_m^2 = \frac{\sigma_\gamma^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \ , \tag{2.16}$$

whereas R_c^2 and R_m^2 correspond to the explained variance of the full models. Furthermore, this ensures a proper decomposition of the full model's R_c^2. This can be easily seen by rearranging equation (2.16) as

$$R_c^2 = \frac{\sigma_\gamma^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} + R_m^2 \ .$$

2.5 Tree-based methods

2.5.1 General theory of tree-based methods

We will begin by introducing regular decision trees and then proceed to introduce bagging and random forest models before describing how variable importance is defined in these models.

Decision trees divide the predictor space into n smaller non-overlapping regions, R_j, where $j \in \{1, \cdots, n\}$. The way predictions are made is by assigning a value to each region, and the observations that fall into region j are predicted to be the average of the response of the training samples that fell into the same region, j.

When creating a decision tree, the goal is to find the boxes $R_1, \cdots, R_n$ that minimize the residual sum of squares(RSS), defined by

$$\sum_{j=1}^{n} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \ ,$$

where $\hat{y}_{R_j}$ is the mean of the response of the observations that fall into region R_j in the training data. A top-down, greedy approach called *recursive binary splitting* is commonly used to partition the predictor space since it is infeasible to consider every possible way of partitioning the predictor space. The approach selects a predictor X_p, $p \in \{1, \cdots, P\}$, where P is the total number of predictors, and a cut-off s for that predictor such that the predictor space is divided into two regions $\{X|X_p < s\}$ and $\{X|X_p \geq s\}$. The values for p and s are chosen such that it gives the biggest decrease to the RSS. That is equivalent to minimizing the expression

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \ ,$$

where $R_1(p, s) = \{X|X_p < s\}$ and $R_2(p, s) = \{X|X_p \geq s\}$ are the potential new regions after the split, and $\hat{y}_{R_1}$ is the mean of the responses for the observations that fall into region 1 in the training set and $\hat{y}_{R_2}$ is the mean of the responses for

the observations that fall into region 2. This process is repeated on one of the two new regions obtained from the previous split, resulting in three regions. The process is repeated until a stopping criterion is fulfilled. A stopping criterion can be the requirement that no regions should contain more than five observations.

Once the predictor space is partitioned into the regions $R_1, \cdots, R_J$, each region is assigned a value equal to the average of the response values of the training observations that fall into the respective region. This value will be the predicted response of observations that fall into the region.

Trees created using recursive binary splitting often overfits the training data. Removing some splits from the trees is a well-known strategy to combat this issue. This method is called *pruning*. Removing splits in the tree randomly would not result in very good trees. Intuitively, we want to remove the splits in the tree that give the best results on the test data, that is, least residual squared if the problem is a regression problem. However, it is infeasible to evaluate every possible subtree since that would be extremely computationally heavy. A common way to prune trees is with cost complexity pruning. This method does not consider every possible subtree, T of the full tree, T_0. Instead, it considers a sequence of trees indexed by a tuning parameter, α, which is non-negative. If the tuning parameter, α, is zero, then no pruning will be performed. The pruned tree, $T \subset T_0$, is created such that

$$\sum_{j=1}^{|T|} \sum_{i:x_i \in R_j} (y_i - \hat{y}_{R_j})^2 + \alpha |T|$$

is minimized. Here $|T|$ is the number of terminal nodes of the subtree T.

Regular decision trees often suffer from high variance. This means that if we split the training sets into two parts randomly and proceed to fit a decision tree on both halves, then the resulting trees may look very different. *Bootstrap aggregation*, often referred to as *bagging*, and *random forest* are two different ways of creating trees with lower variance.

If we have a set of training observations, $(X_1, \cdots, X_n)$, that all have the same variance, σ^2, then the mean of the training observations

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ ,$$

has variance $\frac{\sigma^2}{n}$. Hence averaging over a set of observations reduces variance. To reduce the variance of a tree-based method, it is possible to create many trees and then take the mean of the created predictions. This will, in many cases, also lead to an increase in prediction accuracy. However, creating a large number of trees also requires a lot of training data.

Since the distribution of the training data, let's call it D, is unknown, we do not have any distribution to sample more data from. However, we can create a new set of training data D^* by randomly drawing observations from the original training data D and

add the observation to D^*. It's important to note that after an observation is drawn and added to D^* it is not removed from D. This means that D^* can and will most likely contain multiple of one observation. The distribution of D^* converges to that of D as the number of sampled observations increases. This method of upsampling data is called *bootstrapping*. Both bagging and random forest make use of bootstrapping to aggregate more training data.

Bagging is a statistical learning method that is designed to reduce variance and avoid overfitting. It is done by creating B different bootstrapped training sets, $(D_1^*, \cdots, D_B^*)$, and then fit a tree on each training set, which results in B trees, $(\hat{f}_{D_1^*}(x), \cdots, \hat{f}_{D_B^*}(x))$. These B trees produce B different predictions, which can be averaged to obtain one single low variance prediction. The expression for this is

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{D_B^*}(x) \ .$$

Unlike when creating regular decision trees, when bagging is performed the trees are not pruned. Thus they have high variance and low bias. It is desirable that the trees are deep since the variance becomes smaller when averaging the predictions from all B trees.

If there are one or several very important predictors, then many of the B trees will begin with a split on one of those important predictors. This causes the trees to become similar and predictions from the trees to be correlated. Averaging correlated predictions will lead to a lower decrease in variance than if the predictions were uncorrelated.

Random forest is a very similar method to bagging, in the sense that many trees are grown on bootstrapped training samples, then one prediction is made by averaging all the predictions made by each individual tree. The main difference between the two methods is in how splits in the trees are done. In bagging the split can be done on any predictor, and the one that gives the greatest decrease in RSS is chosen. On the other hand, in a random forest, the split can only be done on a selection of m predictors. These m predictors are chosen at random for each new split. Trees created using random forest will often not have the option of splitting on the important variable early, resulting in more variety in the trees. This leads to less correlated predictions, which again leads to a greater decrease in the variance when averaging the predictions made by all the trees.

The main difference between bagging and random forests is the number of predictors, m, in the set of possible predictors to make a split on. In the `RandomForest` function in the `RandomForest` package in `R` the default number of predictors, m, is set to be $\sqrt{p}$ for classification problems and $\left\lfloor \frac{p}{3} \right\rfloor$ in regression problems. Here $\lfloor . \rfloor$ is the floor-function which rounds down to the nearest integer (e.g $\lfloor 4.8 \rfloor = 4$). Breiman (2001) states that the mean squared errors of the predictions decrease with more features, while correlations also increase, but the correlations increase slow, thus more features are suggested. Liaw et al. (2002) suggests trying either the default, half the default, or double the default for this parameter.

There are different ways of assessing the performance of a bagging model or a random forest model. One approach is to split the data set into training and test set before training the model on the training set and then look at how well it performs on the test set by looking at the residuals squared. Another approach is to use a method called out-of-bag (OOB) error estimation. This method utilizes that on average only $\frac{2}{3}$ of the observations are included in a bootstrapped training sample. This means we can get a prediction for observation i by using the trees where observation i was not used to create the tree. This is referred to as observation i being out-of-bag. Using OOB error estimation eliminates the need to split the data into training and test sets. In the random forest model, there are in particular two parameters that influence the performance of the random forest, the number of predictors to choose from when making a split, *mtry*, and the number of trees created, *ntree*.

2.5.2 Relative importance in bagging and random forest

It is not as easy to interpret bagging and random forest models as it is to interpret simple decision trees, which can be visualized easily. However, it is possible to calculate the variable importances. This can be done using the RSS in the case of regression trees. The variable importance will be the total amount that the RSS is decreased when a split is performed on the given predictor, then averaged over the B trees. If a predictor contributes on average to large decreases in the RSS, it is regarded as important.

The variable importance metric in random forests differs from that of linear models because it is not decomposing a model statistic. The variable importance of linear models always takes values between zero and one, while random forests' variable importance is only limited to positive values. This makes comparing them not so trivial. One possible solution is to rescale the values for the importances for both methods, such that they sum to one. This way it is possible to compare the two ways of assigning importances, but they have no practical meaning after rescaling.

2.5.3 Categorical regressor variables in random forests

Random forests are a statistical learning method that can handle categorical regressor variables. Categorical variables, $Z_i = \{z_1, \cdots, z_K\}$, are handled in random forests by sending a subset of the categories, $Z \subset Z_i$, to the left and the rest of the categories to the right (Cutler et al., 2012).

By encoding a random factor variable (a variable indicating which cluster or individual an observation belongs to) as a categorical variable it is possible to use a random forest model on data set where observations within one cluster are correlated. Since random forest models can provide variable importance, it serves as a good comparison for the extension of the LMG-method, described in Chapter 2.4.

2.6 Simulating data

Simulated data is data that is artificially created rather than being the result of a data collection process. This is a cost-efficient way of obtaining data when testing how a method works or illustrating a new method. When illustrating how new methods work, it is not uncommon to use simulated data as this allow for tuning of the parameters as well as the ability to increase or decrease the size of the data set, such that it is possible to obtain a food understanding of advantages and limitations of a method. Another benefit of synthetic data is that it is easily reproduced.

When creating a suitable data set for random intercept models, there are some properties it should possess. The expected value of the j^{th} response of the i^{th} individual, Y_{ij}, should have expected value

$$\mathbb{E}(Y_{ij}) = \gamma_i + \beta_0 + \beta_1 x_{ij}^{(1)} + \cdots + \beta_p x_{ij}^{(p)} \ ,$$

where γ_i is the individual specific deviance from the global intercept β_0, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is the vector of slope parameters, $\boldsymbol{x}$ is the covariate matrix, and p is the number of predictors. For the purpose of estimating relative importances it is useful to be able to control the covaraince matrix of $\boldsymbol{x}$, $\boldsymbol{\Sigma}$. Here we draw $\boldsymbol{x}$ from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The random intercepts $\gamma_i + \beta_0$ should be independent of $\boldsymbol{x}$ and the errors ε_{ij} and identically distributed (i.i.d.) from a normal distribution with mean β_0 and variance σ_γ^2. The errors ε_{ij} should also be i.i.d. from a normal distribution with mean zero and variance σ_ε^2. Sampling data suitable for a random intercept model can therefore be done by

$$\begin{aligned} &\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(p)}) \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ &\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2) \\ &Y_{ij} = \gamma_i + \beta_0 + \beta_1 x_{ij}^{(1)} + \cdots + \beta_p x_{ij}^{(p)} + \varepsilon_{ij} \ . \end{aligned} \tag{2.17}$$

The covariance matrix $\boldsymbol{\Sigma}$ is chosen with ones on the diagonal, and it has to be symmetric and positive semi-definite, as this is a requirement for it to be a valid covariance matrix.

Chapter 3

Examples

This chapter is devoted to the examples of the methods described in the theory part. First, we will take a closer look at the simulated dataset and see how the methods perform when varying some of the parameters of the data. Then we will proceed to a real-world example where we will look closer at a study of the physical activity among children.

3.1 Simulated data

The simulation was done using `R` version 4.0.3, and the code is provided in the Appendix. Four different models have been fitted to the data: a regular linear regression model, a linear regression model with one random intercept term, and two random forest models. The random factor variable, Z, was excluded from the regular linear model and from one of the random forest models. The resulting relative importances will be presented at the end of this section.

The simulated data is sampled as described in section 2.6. First, four predictors of $\boldsymbol{X}$, $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$, $\boldsymbol{X}^{(3)}$ and $\boldsymbol{X}^{(4)}$, was sampled from a multivariate normal distribution with mean, $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.3 & 0 & 0 \\ -0.3 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.6 \\ 0 & 0 & 0.6 & 1 \end{bmatrix}.$$

We then proceeded to create a random factor variable, Z. This was done using a categorical variable with 40 factors. Each factor was assigned 200 observations of $\boldsymbol{X}$. Due to the limitation of 53 factors in the `randomForest` package, the number of individuals had to be smaller than this number since it would be used as a factor variable. We chose 40 individuals in this case, with 200 observations each.

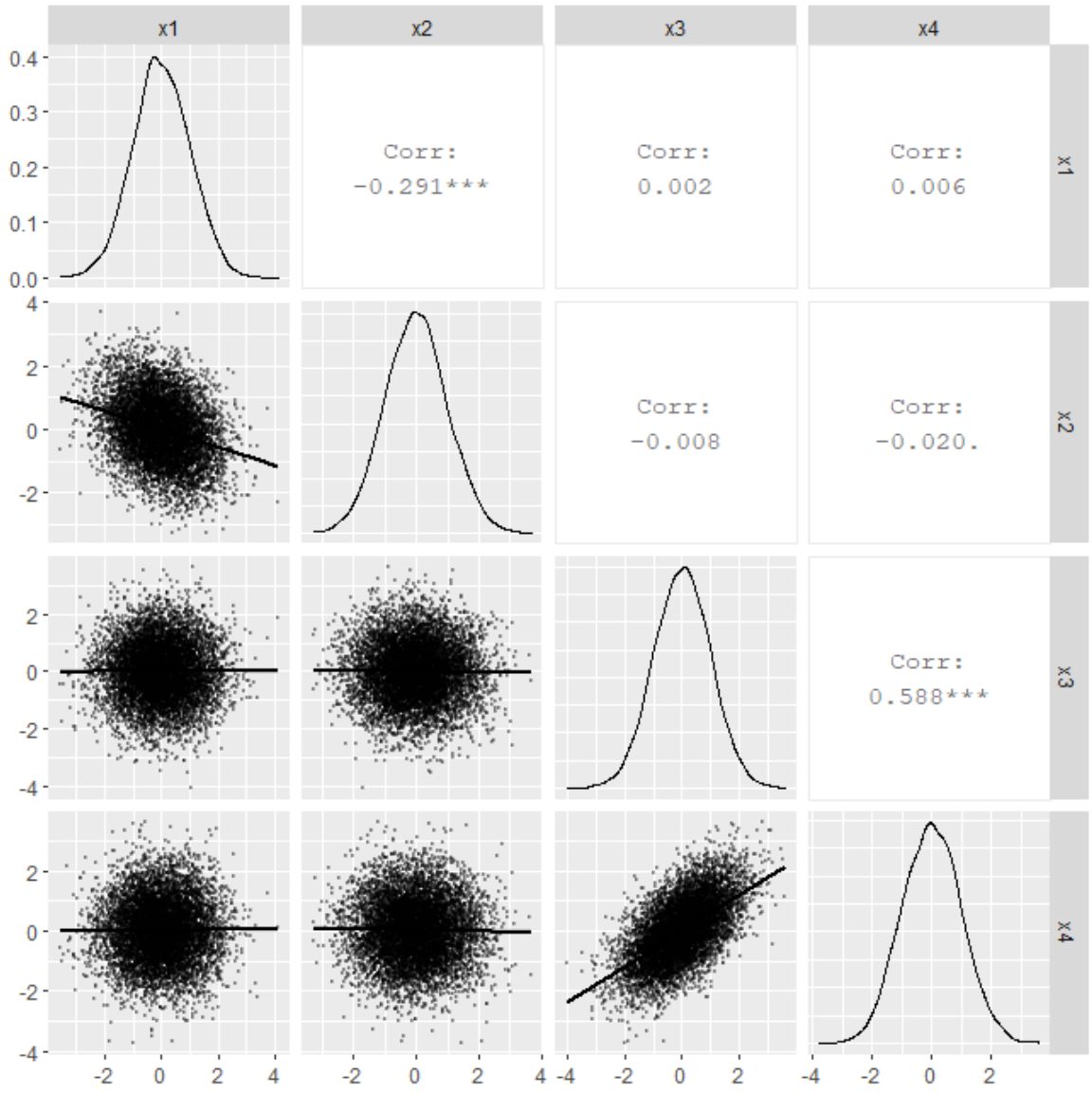

Figure 3.1: Pairs plot of $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$, $\boldsymbol{X}^{(3)}$ and $\boldsymbol{X}^{(4)}$. The diagonal elements show the distribution of the variable coresponding to that column. The upper triangular elements shows the correlation between the variables of the respective row and column. The lower triangular elements shows scatterplots between the variables of the respective row and column.

Figure 3.1 show the pairs plot of the first four regressors in $\boldsymbol{X}$. The figure reflects that $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are negatively correlated with correlation factor close to -0.3 and that $\boldsymbol{X}^{(3)}$ and $\boldsymbol{X}^{(4)}$ are positively correlated with correlation factor close to 0.6. All variables are normally distributed with mean zero.

Next, Y was created according to (2.17) with regression coefficients, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5) = (1, -1, 3, -2, 1)$, error variance $\sigma_\varepsilon^2 = 3$ and variance of the random intercept $\sigma_\gamma^2 = 2$.

We proceeded to fit four different models to the data set, where in two of them the

random factor was ignored and included in the other two. The regular linear model was compared to a random forest model, both models ignoring Z to make it a fair comparison. Then a random intercept model was compared to a random forest model where Z were encoded as a categorical variable.

Random effect	**LMM**		**LM**	
	Variance	SE	Variance	SE
Z	1.805	1.343	-	-
Residuals	3.024	1.739	4.782	2.187
Fixed effect estimates				
	Estimate	SE	Estimate	SE
β_0	1.16	0.213	1.16	0.024
β_1	-1.00	0.020	-0.99	0.025
β_2	2.97	0.021	2.95	0.026
β_3	-2.01	0.024	-2.03	0.030
β_4	0.99	0.024	1.02	0.030

Table 3.1: Coefficient estimates and standard errors(SEs) from fitting a random intercept model(LMM) and a regular linear model without the random intercept(LM) on a simulated data set.

The random intercept model was fitted using the `lmer` function in the `lme4` package (Bates et al., 2015). The summary of the two linear model fits shows that the resulting parameter estimates are very close to the specified values given by the simulation setup (Table 3.1), which is as expected. The fixed population intercept, β_0, has a much larger standard error in the LMM than the LM. The proportion of variance that is explained, the R^2_m and R^2_c statistics, of the random intercept model is 0.751 and 0.844 respectively, and the regular linear model has an R^2 of 0.743. The random factor, Z, was not included in the regular linear model. We could have included it as a factor variable instead, but that would have created issues when calculating the relative variable importance, since the `relaimpo` package is not yet designed to handle factor variables with many levels.

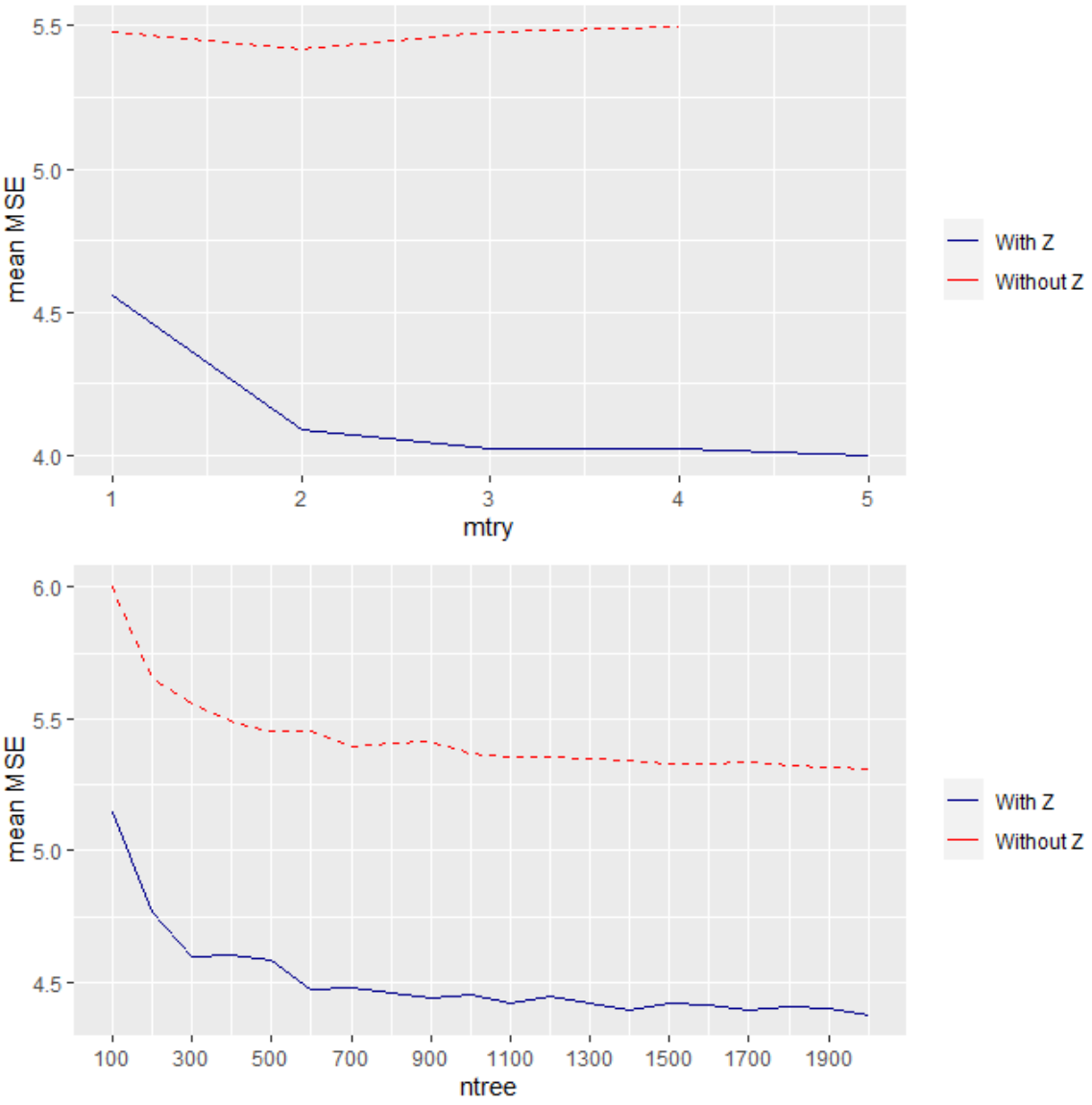

Figure 3.2: The upper plot shows the mean OOB error estimates (MSE) when varying the number of predictors to split on when making the trees in a random forest. The lower plot shows the MSE when varying the number of in the random forests. The solid blue line is from the model where Z is a categorical variable, and the red dashed line is from the random forest where Z is excluded.

We will now look at the random forests and see how different choices of parameters affected the performance of the trees. We wanted to choose values of parameters *mtry* and *ntree* that resulted in a low OOB error estimate. The number of regressors in the set of regressors the tree can choose from when making a split, *mtry*, has default value one in this case, since $\lfloor \frac{5}{3} \rfloor = 1$. However, as seen in the left plot in Figure 3.2, choosing $mtry = 2$ resulted in considerably lower OOB error estimates, at least when Z was included in the model. The models used when assessing different values of *mtry* had the default number of trees in the forest, that is 500.

The right plot in 3.2 shows how the number of trees in the forest influences the random forest's performance.

It is well known that more trees improve the model, but at the cost of computational intensity. The change in MSE seems to be small when using more than 1500 trees, with no obvious improvement. Therefore 1500 trees were be used when calculating the importances. *mtry* was chosen to be 2 when calculating the OOB error estimates.

Now that the four different models were created, we wanted to see how the variable importances compared. We therefore proceeded to calculate the relative importances of the variables in the different models. To ease the presentations of the results we will rename the models.

- *Model 1*: The regular linear model where the random factor variable Z is not included.
- *Model 2*: The random forest model where the random factor variable Z is not included.
- *Model 3*: The random intercept model with Z as random intercept variable.
- *Model 4*: The random forest model where Z is encoded as a categorical variable.

Regressor	**Importances M1**		**Importances M2**		**Importances M3**		**Importances M4**	
	Regular	Std.	Regular	Std.	Regular	Std.	Regular	Std.
$\boldsymbol{X}^{(1)}$	0.155	0.155	23942	0.165	0.099	0.118	19616	0.134
$\boldsymbol{X}^{(2)}$	0.483	0.650	82293	0.568	0.454	0.538	74925	0.513
$\boldsymbol{X}^{(3)}$	0.126	0.169	26190	0.181	0.154	0.182	20376	0.140
$\boldsymbol{X}^{(4)}$	0.018	0.025	12417	0.086	0.044	0.052	6908	0.047
Z	-	-	-	-	0.093	0.110	24117	0.165

Table 3.2: The importances assigned to the regressors in the four models, along with the standardized importances. M1 denotes model 1 with equivalent notation for M2, M3 and M4. "Std." stands for standardized, which in this case means that they sum to one.

The scale of the importances from the linear models and the random forest models differed significantly. Therefore, standardized importances were also calculated in order to compare the importances from the different models. An overview of the importances is given in Table 3.2. The importances will be presented in more detail later in this Chapter.

S	$R_m^2(S)$	$R_m^2(S \cup \boldsymbol{X}^{(1)})$	Weight	LMG-summands
$\emptyset$	0	0.184	$\frac{6}{24}$	0.046
$\{\boldsymbol{X}^{(2)}\}$	0.551	0.560	$\frac{2}{24}$	0.001
$\{\boldsymbol{X}^{(3)}\}$	0.111	0.294	$\frac{2}{24}$	0.015
$\{\boldsymbol{X}^{(4)}\}$	0.003	0.186	$\frac{2}{24}$	0.015
$\{\boldsymbol{X}^{(2)}, \boldsymbol{X}^{(3)}\}$	0.657	0.676	$\frac{2}{24}$	0.002
$\{\boldsymbol{X}^{(2)}, \boldsymbol{X}^{(4)}\}$	0.552	0.561	$\frac{2}{24}$	0.001
$\{\boldsymbol{X}^{(3)}, \boldsymbol{X}^{(4)}\}$	0.229	0.396	$\frac{2}{24}$	0.014
$\{\boldsymbol{X}^{(2)}, \boldsymbol{X}^{(3)}, \boldsymbol{X}^{(4)}\}$	0.727	0.750	$\frac{6}{24}$	0.006
Sum			1	0.099

Table 3.3: The contributions to the importance of $\boldsymbol{X}^{(1)}$ from all subset models of the full model. Recall that S is the set of regressors that appear before X_1 in a model. The weight is given by $\frac{n(S)!(p-n(S)-1)!}{p!}$, where p is the number of regressors in S.

In order to get a better understanding of how variable importances are assigned by the extended LMG-method, we will look in more detail at the importance assigned to $\boldsymbol{X}^{(1)}$ in model 3, since this is a newly proposed method. Table 3.3 shows the contribution from all the subset models of the full model. The largest contribution came from the increase in R_m^2 when $\boldsymbol{X}^{(1)}$ was added to the empty model. The importances of $\boldsymbol{X}^{(2)}$, $\boldsymbol{X}^{(3)}$ and $\boldsymbol{X}^{(4)}$ were calculated equivalently. The importance of the random intercept was calculated as the difference between R_c^2 and R_m^2, which resulted in an importance of $0.844 - 0.751 = 0.093$.

It is of interest to compare the importances assigned by the LMG-method for regular linear regression models to those assigned by random forests without categorical variables to get an idea of what to expect when taking the step to random intercept models. The importances of the variables in model 1 were calculated using the LMG-method for regular linear regression models. This can be done using the `relaimpo` package, made by Grömping (2006). Model 2 was created using the `randomForest` package, which let us extract the importances directly (Liaw et al., 2002).

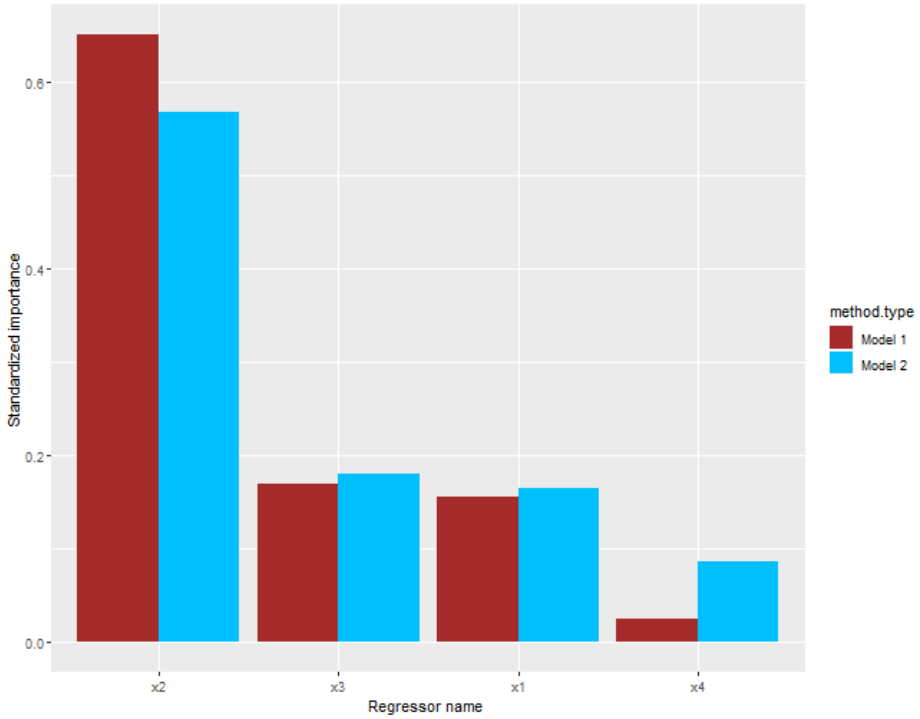

Figure 3.3: The standardized importances of the regular linear model (Model 1, in red) and the random forest model (Model 2, in blue). The random factor, Z, is excluded from both models.

The two methods resulted in quite similar standardized importances for the regressors (Figure 3.3). $\boldsymbol{X}^{(2)}$ was regarded as the most important regressor in both models and $\boldsymbol{X}^{(4)}$ least important. $\boldsymbol{X}^{(4)}$ was regarded as more important relative to the other variables in model 2 than in model 1 as seen in Figure 3.3.

The next step was to compare the variable importances of a random intercept model to a random forest model where the random factor, Z, is encoded as a categorical variable in the random forest approach. We have already seen how the importances of the variables in the random intercept model were calculated in the above example where the importance of $\boldsymbol{X}^{(1)}$ was derived. Furthermore, notice that the regular importances of the fixed effects in model 3 in Table 3.2 sum to the R^2_m of the full model, as expected. It is not that surprising that if you add the importance of Z to R^2_m you get R^2_c, since the random intercept importance is the difference between R^2_c and R^2_m. Model 3 was fitted using the `lme4` package (Bates et al., 2015). The `randomForest` package was used to fit model 4 as this package can also handle categorical variables and provide an importance measure of the variables.

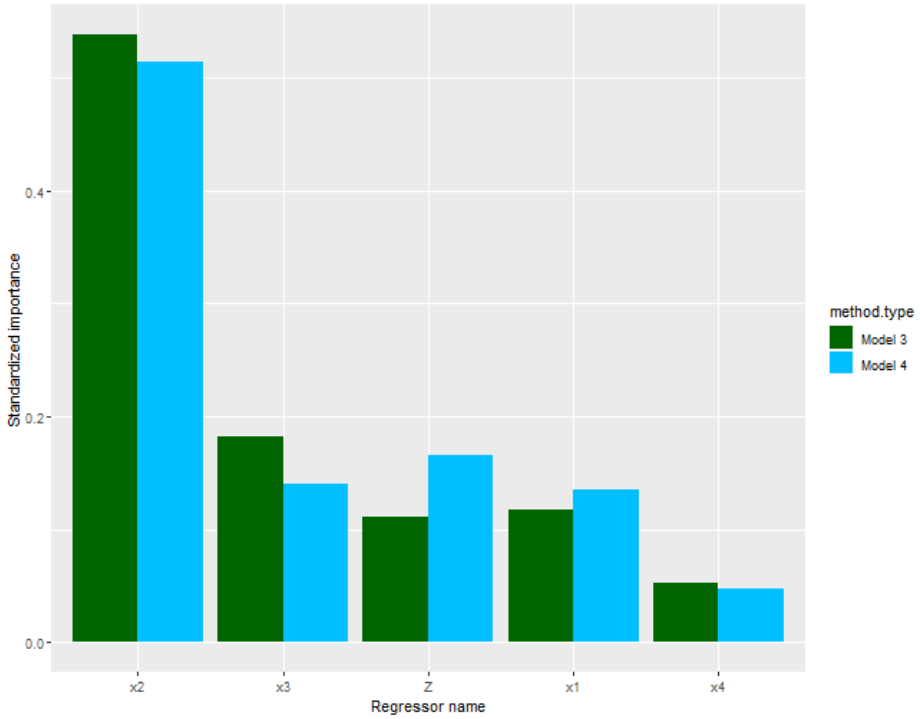

Figure 3.4: The standardized importances of the random intercept model (Model 3, in green) and the random forest model (Model 4, in blue). The random factor, Z, was used as random intercept in the random intercept model and as a categorical variable in the random forest.

The two methods resulted in fairly similar standardized importances(Figure 3.4). Both in model 3 and model 4, $\boldsymbol{X}^{(2)}$ was regarded as the most important variable and $\boldsymbol{X}^{(4)}$ was regarded as the least important variable, just like in the models without the random factor variable Z.

When calculating the relative importances we wanted to see how varying some of the parameters of the simulated data set would influence the importances assigned in the different methods. From the expression for the variance of the fixed effects in equation (2.12) we see that it can be of interest to vary the slope coefficient of one of the predictors and the covariance between two regressors. We used the same names for the renamed models as in the example where the simulation parameters were fixed.

It is expected that the importances in the linear models, model 1 and model 3, to be affected by the variations in the parameters, but it is interesting to see whether the random forest models, model 2 and model 4, behave similarly to the linear models. Therefore, the slope coefficients we used when calculating the relative importances was $(\beta_1, 3, -2, 1)$ with covariance matrix for $\boldsymbol{X}$ equal to

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.3 & 0 & 0 \\ -0.3 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.6 \\ 0 & 0 & 0.6 & 1 \end{bmatrix},$$

where we let β_1 vary from -5 to 5. When varying the covariance coefficient we used slope coefficients $(\beta_1, \beta_2, \beta_3, \beta_4) = (-1, 3, -2, 1)$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \Sigma_{12} & 0 & 0 \\ \Sigma_{12} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.6 \\ 0 & 0 & 0.6 & 1 \end{bmatrix}.$$

and let Σ_{12} vary from -0.9 to 0.9 since these are values that are found in a typical regression model.

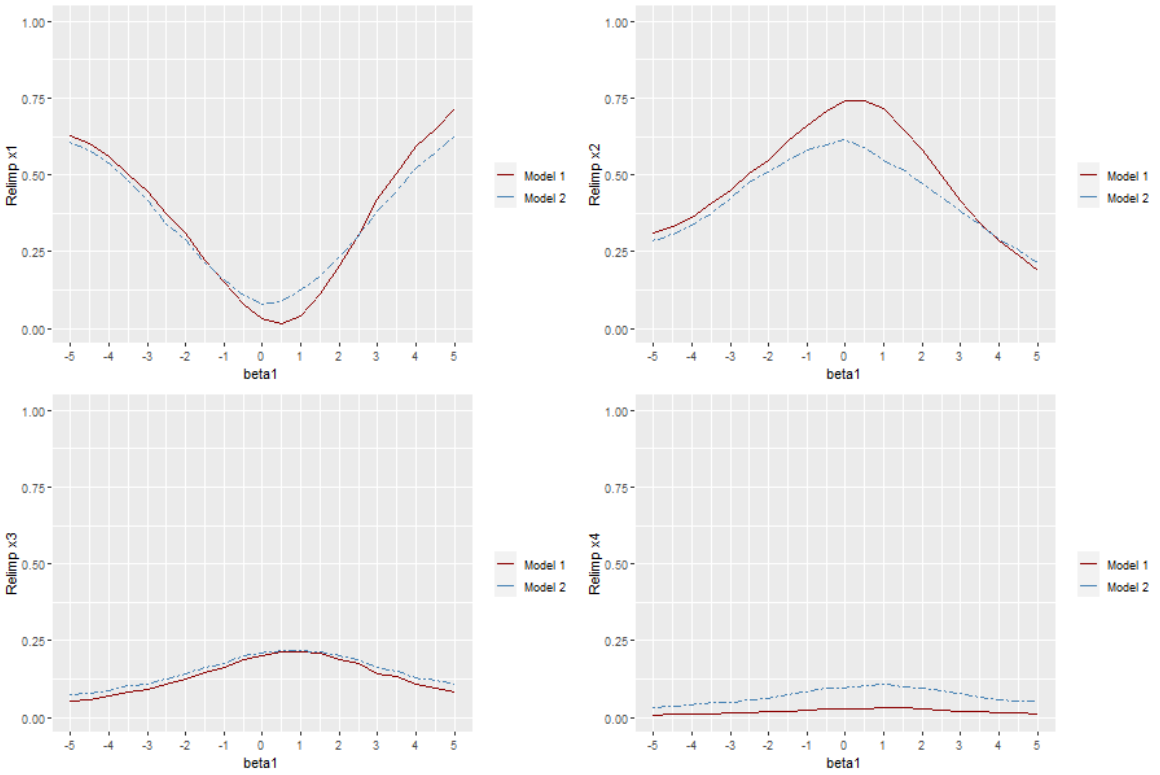

Figure 3.5: The standardized variable importances assigned to the variables in model 1 and model 2 when varying the slope coefficient, β_1, in the simulated data. The red line shows the importance assigned in a regular linear regression model (model 1) and the blue line shows the importance assigned in a random forest model (model 2).

The importances assigned to the variables in model 1 were influenced by the change in β_1 as expected. The standardized importance of the corresponding regressor, $\boldsymbol{X}^{(1)}$, was low when the absolute value of β_1 was low, and vice versa. The standardized importances in model 2 followed a quite similar slope to the standardized importances

of model 1 (Figure3.5). However, the importances in model 2 appeared to be slightly less affected by the change in β_1. Furthermore, we see that the importance of $\boldsymbol{X}^{(1)}$ was mainly absorbed by $\boldsymbol{X}^{(2)}$, and a small portion was absorbed by $\boldsymbol{X}^{(3)}$.

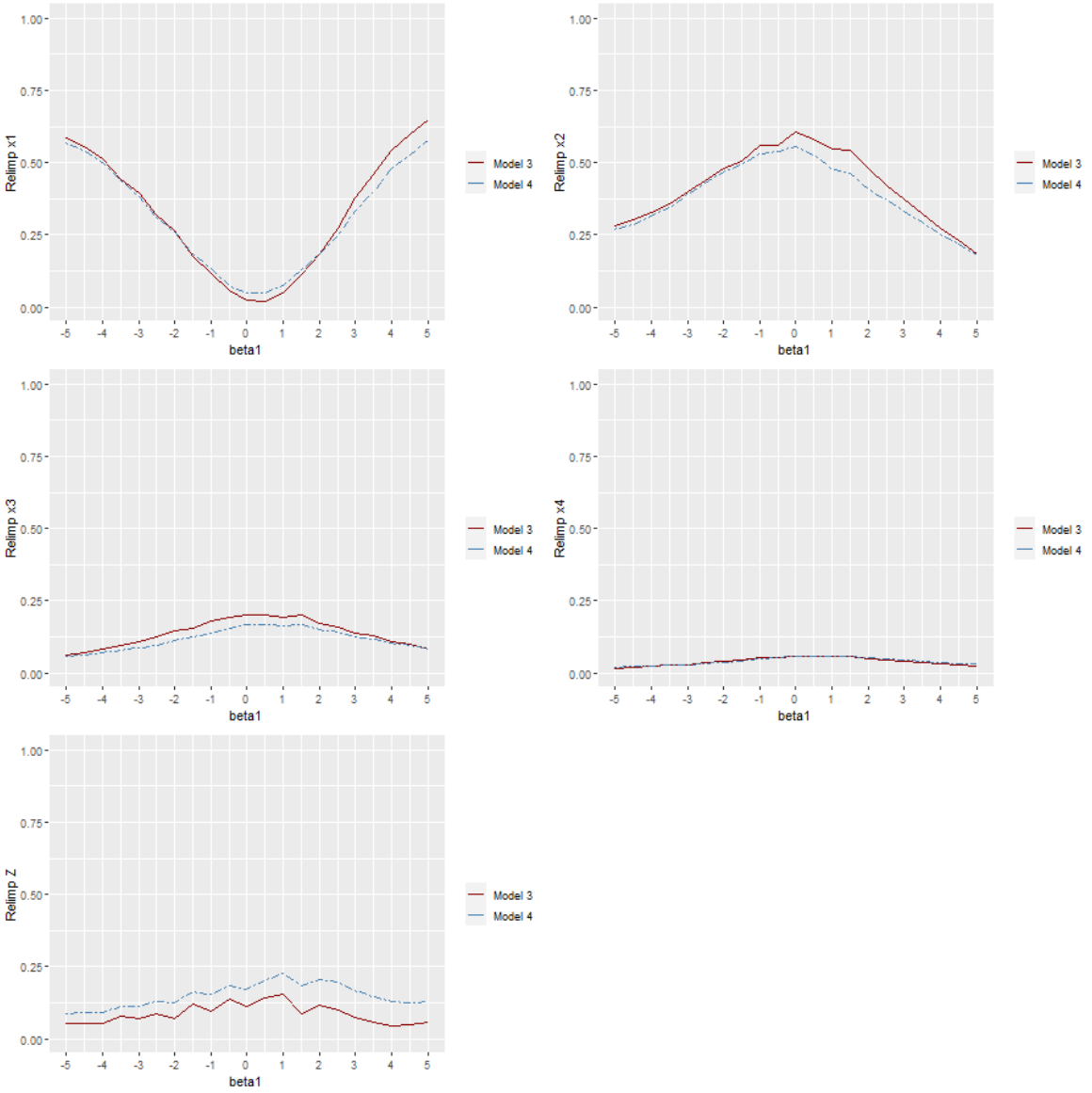

Figure 3.6: The standardized variable importances assigned to the variables in model 3 and model 4 when varying the slope coefficient, β_1, in the simulated data. The red line shows the importance assigned in a random intercept model (model 3) and the blue line shows the importance assigned in a random forest model (model 4).

The standardized importances assigned in the two models, model 3 and model 4, matched quite well when varying the slope parameter of $\boldsymbol{X}^{(1)}$, β_1 (Figure 3.6). The random forest model (model 4) assigned a higher importance to the random factor variable, Z, for the whole interval of β_1. In model 4 the regressors $\boldsymbol{X}^{(2)}$ and $\boldsymbol{X}^{(3)}$ were assigned lower importances than in model 3 for some intervals of β_1, this may be caused by the high importance assigned to Z since the importances are standardized,

but it may also not be the case. The importance of $\boldsymbol{X}^{(1)}$ was mainly absorbed by $\boldsymbol{X}^{(2)}$ and $\boldsymbol{X}^{(3)}$.

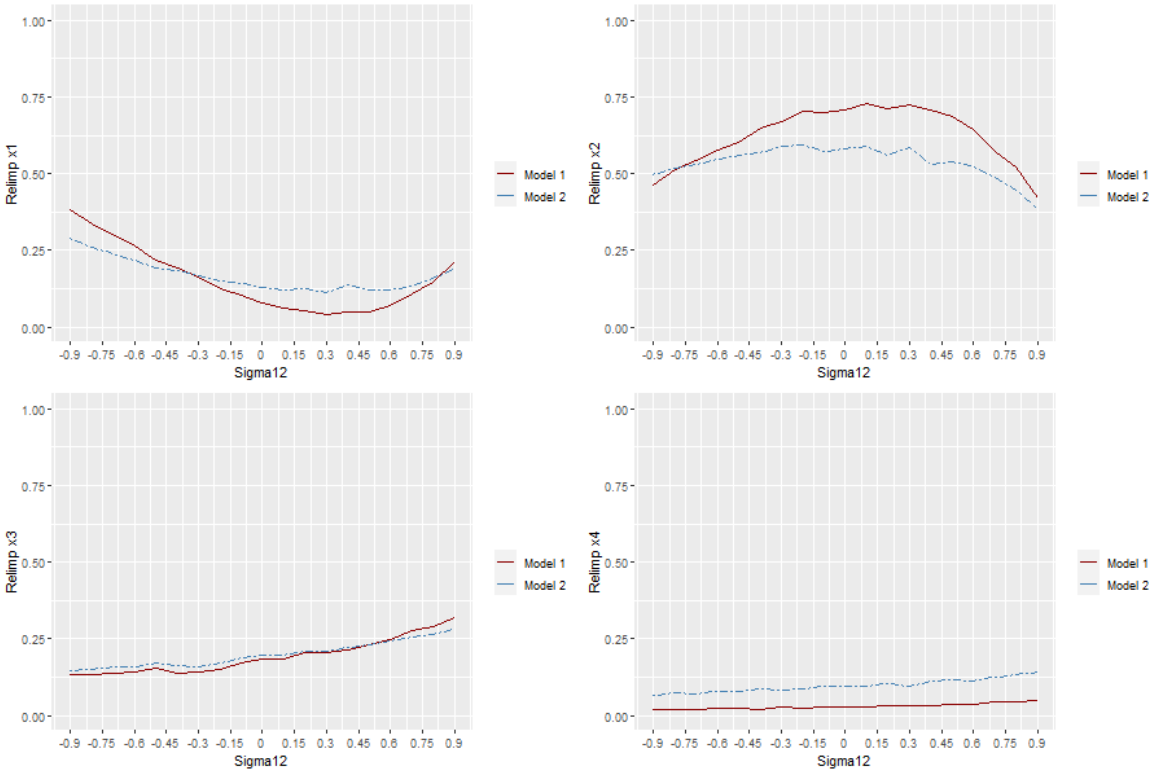

Figure 3.7: The standardized variable importances assigned to the variables in model 1 and model 2 when varying the covariance between the regressors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, Σ_{12}, in the simulated data. The red line shows the importance assigned in a regular linear regression model (model 1) and the blue line shows the importance assigned in a random forest model (model 2).

The importances assigned to the variables in model 1 were influenced by the change in the covariance between the regressors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, Σ_{12}, as expected (Figure 3.7). The standardized importances of the variables in the random forest model (model 2) appear to be less affected by the change in Σ_{12} as seen in Figure 3.7. The importance of $\boldsymbol{X}^{(1)}$ was lowest when Σ_{12} is 0.3. The importance seems to mainly be absorbed by $\boldsymbol{X}^{(2)}$. Also, the importance of $\boldsymbol{X}^{(3)}$ increased as Σ_{12} increased.

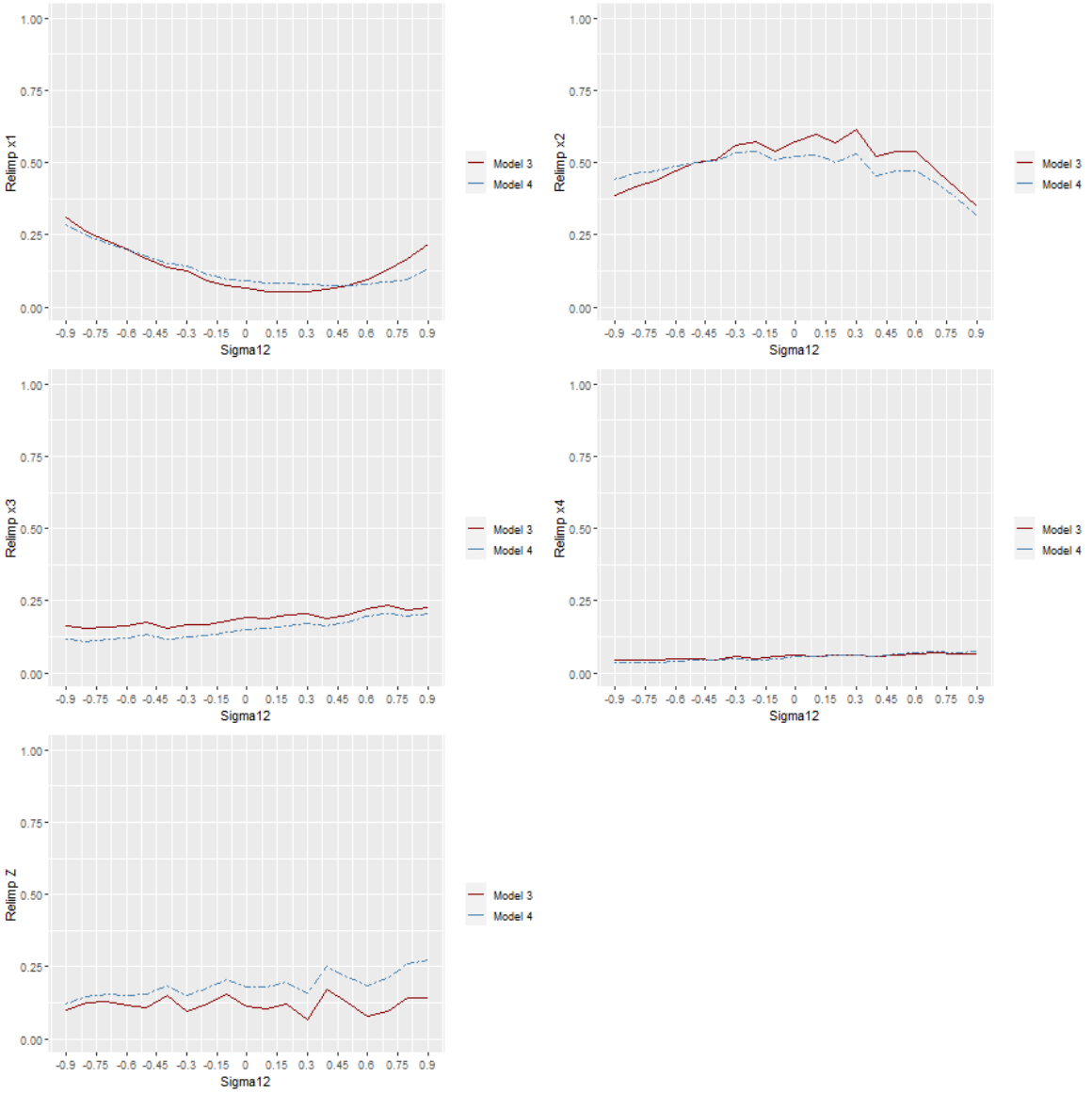

Figure 3.8: The standardized variable importances assigned to the variables in model 3 and model 4 when varying the covariance between the regressors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, Σ_{12}, in the simulated data. The red line shows the importance assigned in a random intercept model (model 3) and the blue line shows the importance assigned in a random forest model (model 4).

The standardized importances of model 3 and model 4 matched quite well when varying the covariance between the regressors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, Σ_{12} (Figure 3.8). The importance of the random factor, Z, was higher in the random forest model (model 4) than in the random intercept model (model 3) as seen in Figure 3.8. It seems like the importances of the variables in model 4 were slightly less affected by the change in Σ_{12}, than the importances in model 3. It also appears like $\boldsymbol{X}^{(2)}$ absorbed the importance of $\boldsymbol{X}^{(1)}$ when Σ_{12} was around 0.3.

3.2 The SPLASHY data

The data set that will be analyzed in this chapter is the Swiss Preschoolers' health study (SPLASHY) data set containing information about children's activity level from different child care centers (Messerli-Bürgy et al., 2016). It contains observations of 476 children across 84 childcare centers located in five cantons of Switzerland (Aargau, Bern, Fribourg, Vaud, and Zurich; Schmutz et al., 2017). The aim of the study was to identify correlates of physical activity. They used a random intercept model to assess the association between the 35 correlates and the total physical activity (TPA).

Schmutz et al. (2017) provide estimates of the relative variable importances using the `relaimpo` package in `R` (Grömping, 2006). Since the `relaimpo` package estimates importances using the LMG-method for a regular linear regression model, Schmutz et al. (2017) only derive relative importance values for a model that disregards the random intercept for childcare. In this Chapter, the newly proposed method of estimating relative variable importance in random intercept models, described in chapter 2.4, will be implemented to see how much the assigned importances differ. Two random forest models were also fitted to the data set, one where the childcare variable was disregarded, and one where the childcare variable is encoded as a categorical variable. The resulting importances from this model will also be provided.

In the SPLASHY data set, the children were between two and six years, of them 54% were boys. The study measured 35 potential correlates, however, Schmutz et al. (2017) only selected the 13 variables with low enough p-values when using a regular linear regression model. The same 13 regressors will be used in this thesis in addition to the random intercept, which will represent the childcare center. For the purpose of fitting the decision trees on the data, only the child care centers that had four or more children in them were part of the analysis in this thesis, since random forests created using the `R` package `randomForest` is limited to factor variables with 53 or fewer categories (Liaw and Wiener, 2002). The data set we were left with after filtering contained observations of 275 children across 43 childcare centers.

Regressor	**Variable type**	**Description**
Sex	Binary	Boys are encoded as one and girls as 0
Age	Continuous	The age of the child at the beginning of the study
Birth weight	Continuous	The birthweight of the child in 100 grams
Gross motor skills	Continuous	The child's score on the Zurich Neuromotor Assessment 3-5
Siblings	Binary	The presence of older siblings in the household is encoded as 1
Family structure	Binary	Single parent household is encoded as 1, dual as 0
Activity temperament	Continuous	The child's mean rating on activity from the EAS Temperament survey. Between 1 and 5
Sport club	Binary	If at least one of the child's parents has sports club membership, it is encoded as 1
Alcohol consumption	Binary	If at least one parent consumes large amounts of alcohol, it is encoded as 1
Time outdoors	Continuous	The number of hours a child spend outdoors per day
Fixed toys	Continuous	Number of fixed play items in the home environment. (E.g. Trampoline)
Neighborhood safety	Continuous	A score of the neighborhood safety between 0 and 44
Season	Factor	Season established using the date of accelerometer
Childcare Center	Factor	The childcare center at which the child was located

Table 3.4: The 13 regressor variables from the physical activity study performed by Schmutz et al. (2017) with descriptions.

Table 3.4 provides a short description of the considered 13 regressors used in the

analysis presented here. More detailed descriptions can be found in the articles by Messerli-Bürgy et al. (2016) and Schmutz et al. (2017).

Random effect	**LMM**		**LM**	
	Variance	Standard error	Variance	Standard error
Childcare center	3113	55.79	-	-
Residuals	15455	124.32	18056	134.37
Fixed effect estimates				
	Estimate	SE	Estimate	SE
Intercept	61.11	23.037	4.16	13.899
Sex	31.46	2.631	33.82	2.614
Age	57.79	1.954	68.01	1.803
Birth weight	1.86	0.235	2.50	0.232
Gross motor skills	18.64	1.351	16.74	1.264
Siblings	-6.88	2.774	-13.30	2.687
Family structure	60.38	4.240	55.85	4.133
Activity temperament	43.67	1.955	45.81	1.895
Sport club	27.53	2.943	26.26	2.853
Alcohol consumption	-57.68	6.109	-45.71	6.119
Time outdoors	11.10	0.934	11.71	0.905
Fixed toys	20.01	0.967	17.74	0.898
Neighborhood safety	-0.38	0.194	-1.16	0.190
Season	39.63	20.364	42.02	3.163

Table 3.5: Coefficient estimates and standard errors(SEs) from fitting a random intercept model(LMM) and a regular linear model without the random intercept(LM) on a subset of the SPLASHY data set.

A summary of the coefficient estimates obtained when fitting the random intercept model, with childcare as random intercept, and the regular linear model, where the childcare variable was excluded, on a subset of the SPLASHY data set is shown in Table 3.5. The left column shows the coefficient estimates obtained when fitting a random intercept model and the results for the regular linear model are given on the right. The two methods result in similar estimates for the fixed effect estimates, although the intercept estimate differs noticeably. The regular linear model has an R^2 of 0.288, and the random intercept model has an R_c^2 of 0.361 and R_m^2 of 0.233.

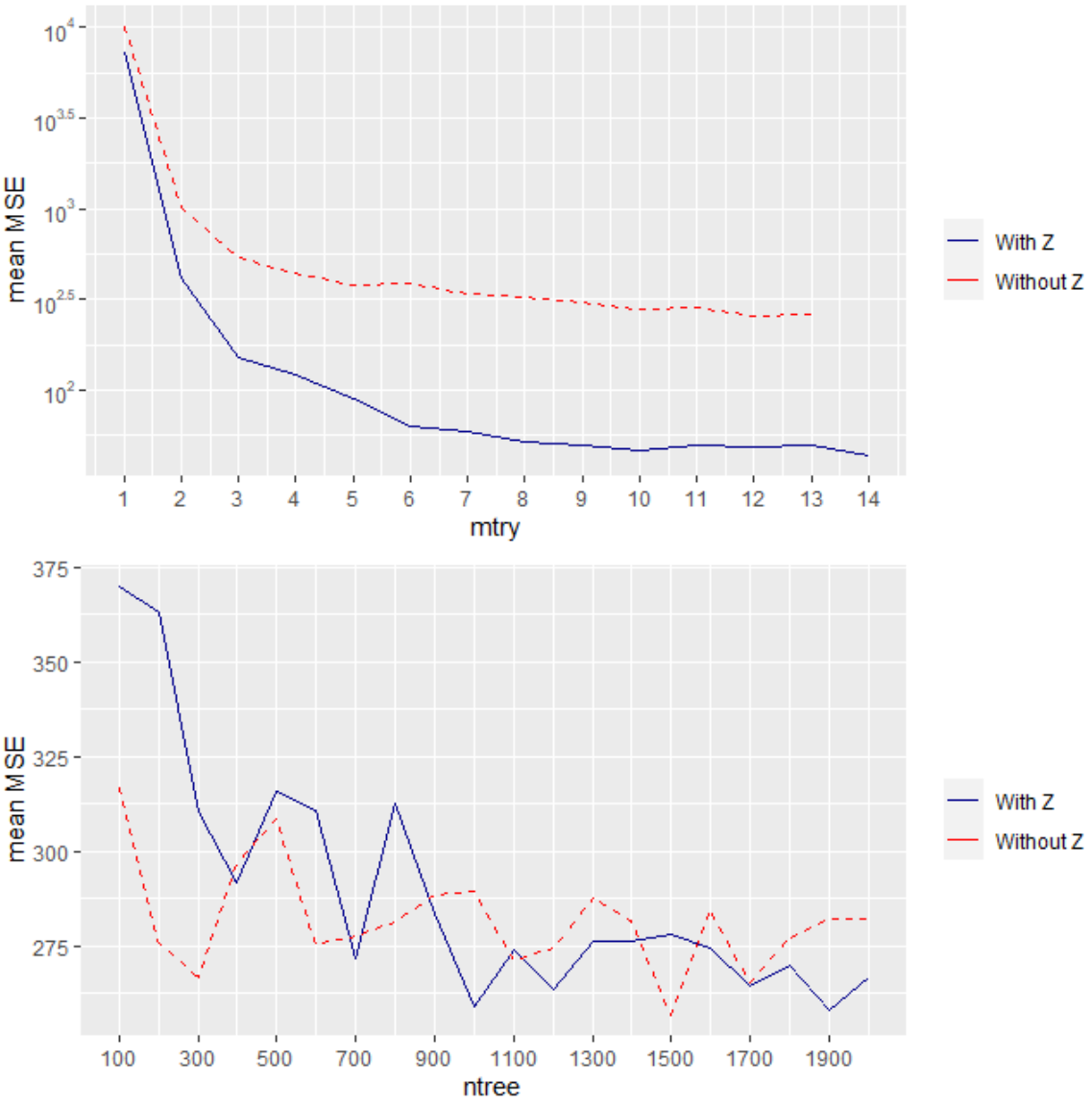

Figure 3.9: The left plot shows the mean OOB error estimates (MSE), in logarithmic scale, when varying the number of predictors to split on when making the trees in a random forest. The right plot shows the MSE when varying the number of in the random forests. The solid blue line is from the model where childcare center is a categorical variable, and the red dashed line is from the random forest where the random factor variable is excluded.

For the purpose of comparing the relative importances, two random forest models were also fitted on the SPLASHY data set using the `R` package `RandomForest`. In Figure 3.9 we see that in general larger *mtry* resulted in better model performance, 10 predictors will be used when calculating the relative importances. The OOB error estimate shows no clear decrease when the number of trees in the model exceeds 1000. Therefore we will use 1000 trees when calculating the relative importances.

Now that the four models were fitted we wanted to see how the importances compared.

To ease the presentations of the results we will rename the models.

- *Model 1*: The regular linear model where the random factor variable, childcare center, is not included.
- *Model 2*: The random forest model where the random factor variable, childcare center, is not included.
- *Model 3*: The random intercept model with childcare center as random intercept variable.
- *Model 4*: The random forest model where the childcare center variable is encoded as a categorical variable.

Regressor	**Importance M1**	**Importance M3**
Sex	0.017	0.010
Age	0.098	0.077
Birth weight	0.010	0.005
Gross motor skills	0.012	0.016
Siblings	0.007	0.003
Family structure	0.010	0.015
Activity temperament	0.057	0.041
Sport club	0.006	0.007
Alcohol consumption	0.003	0.006
Time outdoors	0.011	0.011
Fixed toys	0.039	0.036
Neighborhood safety	0.009	0.001
Season	0.006	0.005
Childcare center	-	0.129
Sum	0.286	0.233

Table 3.6: The importances assigned to the regressors in model 1 (M1) in the left column and in model 3(M3) in the right column. Model 1 had an $R^2 = 0.286$ and model 3 had $R^2_m = 0.233$ and $R^2_c = 0.361$.

The importances of model 1 were calculated using the package `relaimpo` Grömping (2006). The resulting importances of the variables in model 1 are presented in the left column of Table 3.6. The importances of model 3 were calculated using the `relimpLMM` package which is provided in Appendix A. The importances of the variables in model 3 are shown in the right column of Table 3.6.

Regressor	**Std. Imp. M1**	**Std. Imp. M2**	**Std. Imp. M3**	**Std. Imp. M4**
Sex	0.058	0.020	0.028	0.023
Age	0.343	0.290	0.214	0.167
Birth weight	0.037	0.120	0.014	0.075
Gross motor skills	0.044	0.161	0.043	0.110
Siblings	0.025	0.015	0.007	0.008
Family structure	0.036	0.015	0.041	0.006
Activity temperament	0.199	0.124	0.114	0.072
Sport club	0.021	0.013	0.021	0.010
Alcohol consumption	0.012	0.003	0.016	0.002
Time outdoors	0.039	0.059	0.030	0.035
Fixed toys	0.136	0.076	0.100	0.047
Neighborhood safety	0.030	0.088	0.003	0.056
Season	0.020	0.015	0.013	0.003
Childcare center	-	-	0.356	0.385

Table 3.7: The standardized importances of all four models. M1, M2, M3 and M4 stands for model 1, model 2, model 3 and model 4 respectively. Std. Imp. stands for standardized importance.

The importances of the random forest models, model 2 and model 4, can be extracted directly from the models which were fitted using the `RandomForest` package. To be able to compare the importances of the four models they had to be standardized. Table 3.7 gives an overview of the resulting standardized importances.

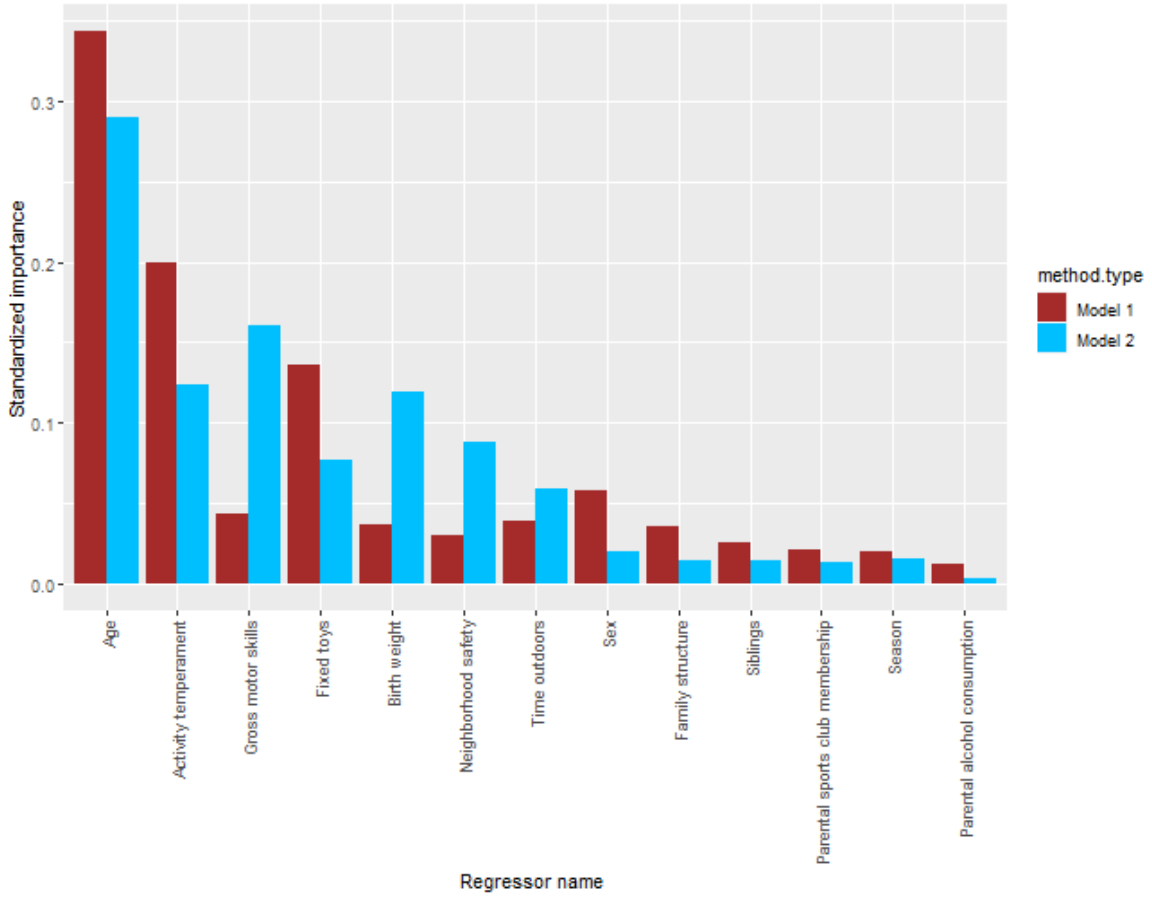

Figure 3.10: Relative importances of the regressors in the SPLASHY data set using the LMG-method for a regular linear model (model 1, in red) and a random forest model (model 2, in blue). The childcare variable is excluded from both models.

Figure 3.10 shows the comparison between the standardized variable importances in model 1 and model 2. In both models, the age variable received a large share of the importance, while Alcohol consumption received a very small share of the importance. On the other hand, gross motor skills and birth weight received relatively unequal shares of importance in the two different models.

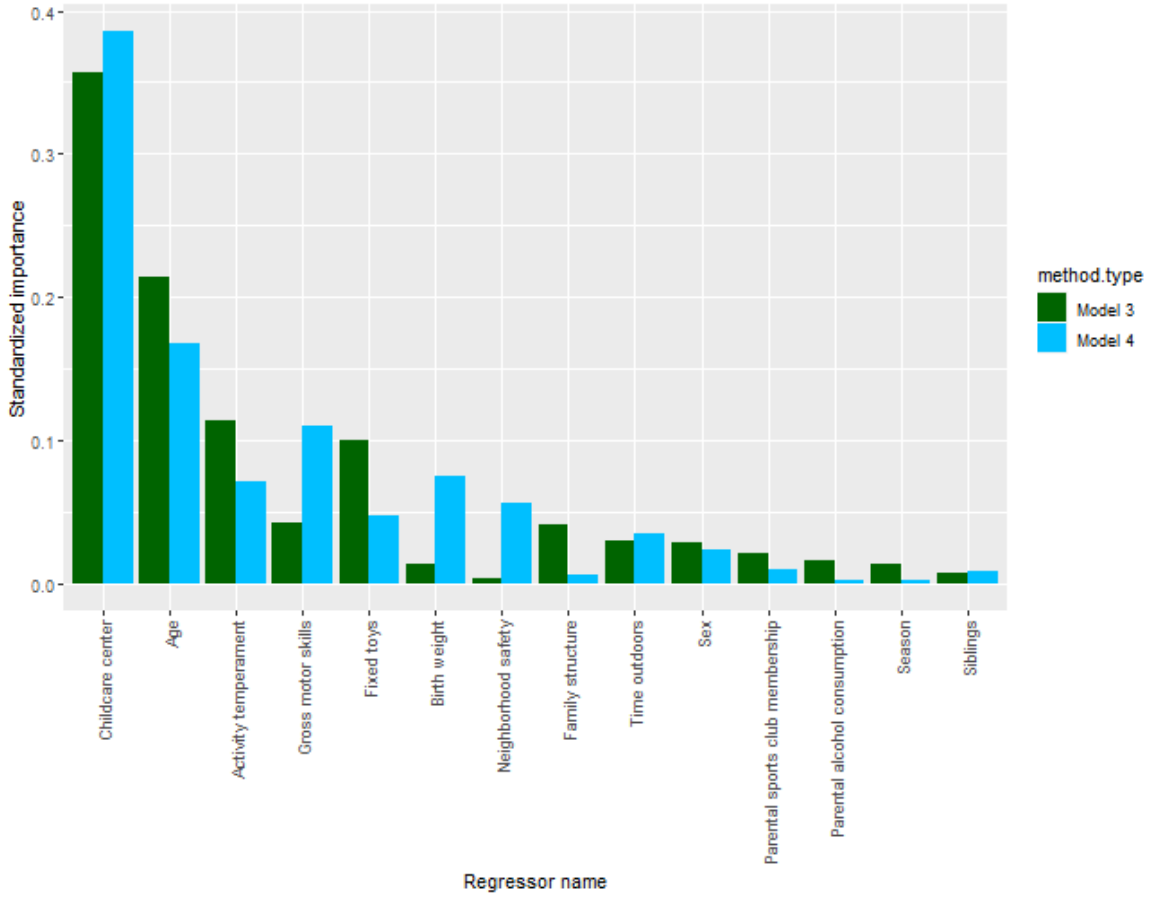

Figure 3.11: Standardized relative importances of the regressors in model 3 (in green) and model 4 (in blue).

Figure 3.10 shows the comparison of the variable importances assigned by the extension of the LMG-method applied on the random intercept model (model 3) and the relative importances assigned to the variable in the random forest (model 4). Both methods assigned a large share of importance to the random factor variable, childcare center.

Chapter 4

Discussion and conclusion

We have looked at ways of extending a method for assigning relative variable importance from regular linear regression models to random intercept models. Byhring (2020) found that not permuting the random intercept and decomposing the R_m^2 resulted in the most realistic shares, and therefore this extension of the LMG-method was used throughout this thesis. To assess this extension of the LMG-method, we wanted to use the variable importances assigned by random forests as a comparison. One of the challenges addressed in this thesis was that random forests cannot directly handle random factor variables, however it is possible to encode the random intercept variable as a categorical variable with one factor for each individual or cluster. The statistical models were implemented, and relative importance was calculated on two data sets: One simulated data set and one real-world example.

In the example with the simulated data set we first presented a simple scenario where the parameters of the sampled data set were fixed with slope parameters $\boldsymbol{\beta}$, and we only had correlation between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, and $\boldsymbol{X}^{(3)}$ and $\boldsymbol{X}^{(4)}$. Later, we looked at how varying some of the parameters of the sampled data set influenced the importances assigned to the variables in the different models.

In the case of fixed parameters, we found that both the random intercept model (model 3) and the random forest model where Z was encoded as a categorical variable agreed on which variable was the most important. The random factor variable, Z, received a higher standardized importance in the random forest model than in the random intercept model.

When varying the slope parameter, β_1, we found that the standardized importance of the corresponding regressor was lowest when the absolute value of β_1 was small in absolute value. The standardized importance of some of the other regressors peaked when β_1 was close to zero, this may be because they absorbed some of the importance from the regressor corresponding to β_1, however it may also be that the proportion of variance that is explained, R^2, of the full model is smaller when β_1 is near zero. In general, the importances assigned by the two methods matched quite well.

When varying the covariance between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, Σ_{12}, we found that the importance assigned to the variables in the random forest model where Z was excluded

(model 2) appeared to be less affected by changes in Σ_{12}, than the importances assigned to the variables in the regular linear regression model (model 1). The importances assigned to the variables in the random forest model where Z was included as a categorical variable (model 4) matched better to those assigned in the random intercept model (model 3). However, the random factor variable Z was assigned a higher importance in model 4 than in model 3.

In the example with the splashy data set we found that both the random intercept model and the random forest model where childcare were used as a categorical variable regarded childcare as the most important variable. There are some variables that get assigned quite different importances in the two methods, in particular birth weight and neighborhood safety. However, we see the same differences appear between the importances in the regular linear model where the childcare variable is excluded and the random forest model where the childcare variable is excluded.

In general, we see that the importances assigned in the random intercept models match well with the importances assigned in the random forest models where the random factor variable is encoded as a categorical variable. In some cases we see that the importance assigned to the random factor variable is higher in the random forest model where the random factor variable is encoded as a categorical variable than in the random intercept models. However, it has been discussed in previous literature that using categorical variables in random forest models may lead to "unfair" importances. It is known that the importance measure tends to favor categorical variables with a larger number of factors (Strobl et al., 2007; Grömping, 2009).

The method presented in this thesis for assigning relative variable importance in random intercept models decomposes the explained variance of the full model into shares explained by each predictor. The method as presented is limited to one random intercept, but could with small adjustments handle several random intercepts. It would also be possible to extend the method to handle random slopes, generalized linear models(GLM) and generalized linear mixed models(GLMM) by providing a way to calculate R^2 in these models. The topic of how to calculate R^2 in random slope models, GLMs and GLMMs is discussed in Nakagawa and Schielzeth (2013) and Johnson (2014). Therefore we expect that the extension of the ideas presented here should be possible

An `R` package with functions to calculate relative importance in random intercept models is provided in Appendix A. The function is limited to one random intercept term, and needs at least three fixed effects, and is limited to a maximum of 13. There is room for improvement when it comes to the speed of the functions in the package. For example, when calculating the importances in the simulated data, an obvious improvement would be to only fit the random intercept model with $\boldsymbol{X}^{(1)}$ and Z only once, however since there were four fixed effect variables in the example, the model with only $\boldsymbol{X}^{(1)}$ and Z was fitted three times, one time for each of the other fixed effects when calculating their importance. One possible solution is to fit all possible subset models of the full model and only store the relevant parameters to calculate the R^2_m of the models.

The LMG-method and the extension of the LMG-method discussed in this thesis

are very computationally heavy. Johnson (2000) proposed a method called *relative wheights* which is much less computationally demanding and has been shown to result in relatively equal importance shares as the LMG-method for regular linear models (Grömping, 2015). It could be interesting to see if this method can be generalized to mixed effect models in future studies.

Bibliography

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Berg, J. (2019, 12). Relative variable importance in linear models. Bachelor thesis NTNU.

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

Byhring, O. (2020, 08). Relative variable importance in linear regression models with random intercept term. Project thesis NTNU.

Cançado, L. P. (2018). Determining predictor importance in multilevel models for longitudinal data: An extension of dominance analysis. *Theses and Dissertations. 1978*.

Cutler, A., D. R. Cutler, and J. R. Stevens (2012). Random forests. In *Ensemble machine learning*, pp. 157–175. Springer.

Feldman, B. E. (2005). Relative importance and value. *https://ssrn.com/abstract=2255827*.

Gregorutti, B., B. Michel, and P. Saint-Pierre (2017). Correlation and variable importance in random forests. *Statistics and Computing 27*(3), 659–678.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician 61*(2), 139–147.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician 63*(4), 308–319.

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics 7*(2), 137–152.

Grömping, U. (2006). Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software 17*(1), 1–27.

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

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate behavioral research 35*(1), 1–19.

Johnson, P. C. (2014). Extension of nakagawa & schielzeth's r2glmm to random slopes models. *Methods in ecology and evolution 5*(9), 944–946.

Kruskal, W. and R. Majors (1989). Concepts of relative importance in recent scientific literature. *The American statistician 43*(1), 2.

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News 2*(3), 18–22.

Liaw, A., M. Wiener, et al. (2002). Classification and regression by randomforest. *R news 2*(3), 18–22.

Lindeman, R. H., P. F. Merenda, and R. Z. Gold (1980). *Introduction to bivariate and multivariate analysis*. Glenview, IL: Scott, Foresman.

Liu, Y., B. Zumbo, and A. Wu (2014, 08). Relative importance of predictors in multilevel modeling. *Journal of modern applied statistical methods: JMASM 13*, 2–22.

Messerli-Bürgy, N., T. H. Kakebeeke, A. Arhab, K. Stülb, A. E. Zysset, C. S. Leeger-Aschmann, E. A. Schmutz, F. Fares, A. H. Meyer, S. Munsch, et al. (2016). The swiss preschoolers' health study (splashy): objectives and design of a prospective multi-site cohort study assessing psychological and physiological health in young children. *BMC pediatrics 16*(1), 1–16.

Nakagawa, S. and H. Schielzeth (2013). A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *Methods in ecology and evolution 4*(2), 133–142.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international Tampere conference in statistics, 1987*, pp. 245–260. Department of Mathematical Sciences, University of Tampere.

Schmutz, E. A., C. S. Leeger-Aschmann, T. Radtke, S. Muff, T. H. Kakebeeke, A. E. Zysset, N. Messerli-Bürgy, K. Stülb, A. Arhab, A. H. Meyer, et al. (2017). Correlates of preschool children's objectively measured physical activity and sedentary behavior: a cross-sectional analysis of the splashy study. *International Journal of Behavioral Nutrition and Physical Activity 14*(1), 1.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC bioinformatics 9*(1), 1–11.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics 8*(1), 1–21.

Zhu, R., D. Zeng, and M. R. Kosorok (2015). Reinforcement learning trees. *Journal of the American Statistical Association 110*(512), 1770–1784.

Appendix

Appendix A

The `R` package with functions to calculate the relative variable importances is available on my github and can easely be downloaded using the `devtools` package. This can be done using the following lines of code:

```
> library(devtools)
> install_github("oliverbyhring/relimpLMM")
> library(relimpLMM)
```

This package includes two functions; `calc.R2(LMM.obj, Marginal = TRUE)` and `calc.relimp.lmm(LMM.obj, response.name)`.

The `calc.R2` function takes a random intercept model, created using the `lmer` function in the `lme4`, as an argument and returns the R^2 of the model as proposed by Nakagawa and Schielzeth (2013). If the *Marginal* argument is not specified, then TRUE will be chosen as defaul and the function will return the R^2_m as defined in equation (2.13). If *Marginal* is set to FALSE, then the R^2_c will be returned by the function.

The `calc.relimp.lmm` function takes a random intercept model, created using the `lmer` function in the `lme4`, and the name of the response as a string as arguments and returns the relative variable importance estimates as a data frame. The importances are estimated using the LMG-method as described in section 2.4.

Appendix B

The functions used to sample the covariate matrix, $\boldsymbol{X}$, and the response, $\boldsymbol{Y}$, is provided below. `create.X`

```
> ## Function that creates a matrix of covariates X
> create.X <- function(number.of.cov = 4,
+                      n.individuals = 40,
+                      obs.per.individual = 200,
+                      sigma.mat = NULL){
+   if (is.null(sigma.mat)){
+     A <- matrix(runif(number.of.cov^2)*2-1, ncol=number.of.cov)
+     sigma.mat <- t(A) %*% A
+   }
+   mu.vec = rep(0,number.of.cov)
+   X = MASS::mvrnorm(n.individuals*obs.per.individual,
+                     mu = mu.vec
+                     , Sigma = sigma.mat)
+   pers <- sort(rep(1:n.individuals,obs.per.individual))
+   X<- matrix(c(X,pers), nrow = 1000)
+   return(data.frame(matrix(X, ncol = number.of.cov+1)))
+ }
> ## Function that samples the response for X
> sample.from.y <- function(X = create.X(),
+                           n.individuals = 40,
+                           obs.per.individual=200,
+                           number.of.cov = 4,
+                           residual.var = 3,
+                           random.intercept.var = 2,
+                           beta = NULL,
+                           global.intercept = 1){
+   if(is.null(beta)){
+     a <- -5:5
+     beta <- sample(a,(number.of.cov))
+   }
+   random.intercept <- sort(rep(rnorm(n.individuals,
+                                      mean = 0,
+                                      sd = sqrt(random.intercept.var)),
+                                obs.per.individual))
+   residual.error <- rnorm(n.individuals*obs.per.individual,
+                           mean = 0, sd = sqrt(residual.var))
+   X.mat <- as.matrix(X[1:number.of.cov])
+   Y <- global.intercept + X.mat%*%beta + random.intercept +residual.error
+   return(Y)
+ }
>
```

NTNU

Norwegian University of
Science and Technology