First line of title
second line of title

Master Thesis in Biostatistics (STA495)

by

Name of student
Matriculation number

supervised by

Name of responsible supervisor (with title)
Name of supervisor (with title and affiliation if external)

Zurich, month year

Contents

Preface

Wenns scheisse läuft, läufts scheisse! Kahn O.

Max Muster
June 2018

Chapter 1

Introduction

The objective of this master thesis is to implement the variable importance measure LMG (named after the authors Lindeman, Merenda, and Gold (Grömping, 2007)) in linear models estimated with Bayesian methods. Bayesian methods have gained popularity because they allow to quantify the uncertainty about parameters and they allow to include prior information.

Regression models are popular in many applied research areas (Nimon and Oswald, 2013). These models provide a tool to find an association between a response variable and a set of explanatory variables. The explanatory variables are also called predictors or covariates. Regression parameters provide information to what extent the response variable is expected to change when one predictor changes by one unit, given all other predictors in the model remain the same. Being aware of this last remark is very important for the correct interpretation of the regression parameters. It shows that the parameter value of a predictor is dependent on the other predictors in the model.

Because predictors are often correlated to some degree to each other, it is obviously not an easy task to find the most important predictors in a model. The first question is, what is meant by the importance of a predictor? There is no easy answer to this question and it is depending on the research issue. Grömping (2015) concludes that there may never be a unique definition of variable importance. There exist different metrics to quantify the importance of predictors. These metrics focus on different aspects and with correlated predictors they lead to different conclusions. A summary of variable importance metrics can be found in Grömping (2015).

A distinction should be made between the importance of predictors in regression models that are used to predict future data and in regression models, applied to find an association between predictors and the response variable. In the first case, the aim is only to reduce the error between the predicted values and the observable values. The underlying association between predictors is of minor importance. In the second case, the focus is on the strength of the relationship between the predictors and the response variable. A predictor may explain little of the response variable, given two other correlated predictors are already included in a regression model. However, this predictor that is unimportant from the regression output may be the main cause of the other two predictorss. Therefore, it may be the most important predictor of this regression model (Grömping, 2007).

The causal relationship between the variables is missing in standard regression models. Studying a predictor, given other variables are already included or using models that contain only the predictor itself, provide only some parts of the bigger picture about the predictor in a model. Which are the most useful variable importance metrics is still an open debate. A convincing theoretical basis is still lacking for all of those metrics. Grömping (2015) recommends to use the existing best practices, until a more profound solution will be found. For variance (or generally goodness-of-fit) decomposition based importance, she recommends to use LMG enhanced with joint contributions or dominance analysis (Grömping, 2015).

Chapter 2

Theoretical background

2.1 LMG variable importance metric

The focus of this master thesis is on the LMG variable importance metric. The LMG is a metric that is based on variance decomposition. The total R^2 of a model is decomposed onto the predictors. Marginal and conditional information are incorporated (Grömping, 2015) . The formulas of this section are taken from Grömping (2015), using the same mathematical notations.

The following notations for the explained variance (2.1) and sequentially added variance (2.2) simplify the notation of the LMG formula.

$$\text{evar}(S) = \text{Var}(Y) - \text{Var}(Y \mid X_j, j \in S), \tag{2.1}$$

$$\text{svar}(M \mid S) = \text{evar}(M \cup S) - \text{evar}(S), \tag{2.2}$$

, where S and M denote disjoint sets of predictors.

The LMG formula is given below for the first predictor only. Because of exchangeable predictors, this is no loss of generality. $R^2(S)$ can be written as $\text{evar}(S)/\text{Var}(Y)$.

$$\begin{aligned}
\text{LMG}(1) &= \frac{1}{p!} \sum_{\pi permutation} \text{svar}(\{1\} \mid S_1(\pi)), && (2.3)\\
&= \frac{1}{p!} \sum_{S \subseteq \{2,\ldots,p\}} n(S)!\,(p - n(S) - 1)!\,\text{svar}(\{1\} \mid S) && (2.4)\\
&= \frac{1}{p} \sum_{i=0}^{p-1} \left(\sum_{\substack{S \subseteq \{2,\ldots,p\} \\ n(S)=1}} \text{svar}(\{1\} \mid S) \right) \Big/ \binom{p-1}{i} && (2.5)\\
&= \frac{1}{p} \sum_{i=0}^{p-1} \frac{\sum_{\substack{S \subseteq \{2,\ldots,p\} \\ n(S)=1}} \text{svar}(\{1\} \mid S)}{\binom{p-1}{i}}, && (2.6)
\end{aligned}$$

where $S_1(\pi)$ is the set of predecessors of predictor 1.

The different formula writings help to better understand what the calculation is about in the LMG metric. The R^2 of the model including all predictors is decomposed. In the formula on the top (2.3), the LMG value of predictor 1 is represented as an unweighted average over all orderings of the sequential added variance contribution of predictor 1. The formula in the center (2.4), shows that the calculation can be done more efficiently. The orderings with the

same set of predecessors S are combined into one summand. Instead of $p!$ summands only 2^{p-1} summands need to be calculated. The formula on the bottom (2.5) shows that the LMG metric can also be seen as the unweighted average over average explained variance improvements when adding predictor 1 to a model of size i without predictor 1 (Grömping, 2015). The LMG metric is implemented in the R package `relaimpo` (Grömping, 2006).

Chevan and Sutherland (1991) propose that, instead of only using the variances, an appropriate goodness-of-fit metric can as well be used in the LMG formula. They name their proposal *hierarchical partitioning*. The requirements are simply: an initial measure of fit when no predictor variable is present, a final measure of fit when N predictor variables are present, all intermediate models when various combinations of predictor variables are present. The LMG component of each variable is named *independent component* (I). The sum of the independent components (I) results then in the overall goodness-of-fit metric. The difference between the goodness-of-fit when only the predictor itself is included in the model, compared to its independent component (I), is named the *joint contribution* (J) (Grömping, 2015). Hierarchical partitioning is implemented in the `hier.part` package (Walsh and Nally, 2015). When R^2 is chosen as the goodness-of-fit measure, the LMG values are calculated. The hierarchical partitioning function of `hier.part` is used in this master thesis. The hierarchical partitioning function accepts a data frame with the R^2 values of all submodels as input. Of note, the partitioning function of `hier.part` is only guaranteed to work up to nine predictors and does not work at all for more than twelve predictors.

2.2 Appropriate R^2 definitions in the Bayesian framework

The focus of this master thesis is on the standard linear model. For this model, the most widely used goodness-of-fit metric is R^2. Different formulas for R^2 exist Kvalseth (1985), all leading to the same value when an intercept is included and the model is fitted by maximum likelihood.

Two commonly used R^2 definitions are:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.7}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad i = 1, \ldots, n, \tag{2.8}$$

where $\hat{y}_i == \mathrm{E}(y \mid X_i, \hat{\theta})$. $\hat{\theta}$ is the maximum likelihood estimate of the regression coefficients.

When other estimation methods than maximum likelihood are used, equation (2.7) can be negative and equation (2.8) can be bigger than 1. This is not uncommon in a Bayesian regression setting when samples of the posterior parameter distribution are employed. A model that explains more than 100% of the variance is nonsense. A negative R^2 is also difficult to interpret. A negative R^2 may be interpreted as a fit that is worse than the mean of the data. This can make sense for predictive purposes, e.g. when new data from a test set is predicted by leave-one-out crossvalidation (Alexander *et al.*, 2015). For non predicting purposes, a negative R^2 does not make sense. The aim of the LMG formula is to gain some more information about the possible association between variables. A predictor can not explain less than zero variance in the population. To respect the non-negative share property of the LMG formula, the R^2 of submodels should not decrease when adding predictors. Both classical R^2 definitions seem not to be well suited for the LMG metric in the Bayesian framework.

A more reasonable R^2 definition for the LMG formula in the Bayesian framework can be found by noting that the variance of the linear model can also be written as

$$\mathrm{Var}(y) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta}) + \sigma^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta} + \sigma^2, \tag{2.9}$$

where $\boldsymbol{\beta}^\top = (\beta_1 \dots \beta_p)$ are the regression parameters without the intercept. $\boldsymbol{\Sigma_{XX}}$ is the covariance matrix of the predictors.

By using this variance definition Gelman *et al.* (2017) propose to use

$$R^2_{Gelman} = \frac{\text{Var}(\sum_{i=1}^n \hat{y}_i^s)}{\text{Var}(\sum_{i=1}^n \hat{y}_i^s) + \text{Var}(\sum_{i=1}^n e_i^s)}, \qquad i = 1, \dots, n, \tag{2.10}$$

where $\hat{y}_i^s = \text{E}(y \mid X_i, \theta^s)$ and the vector of errors $e_i^s = y_i - \hat{y}_i^s$ and $\theta^s, s = 1, \dots, S$ are draws from the posterior parameter distribution. The R^2 is then guaranteed to be between 0 and 1. The R^2 can be interpreted as a data-based estimate of the proportion of variance explained for new data under the assumption that the predictors are held fixed (Gelman *et al.*, 2017).

In the Bayesian framework, the σ^2 parameter is explicitly modeled in the standard linear regression setting. Therefore, it is possible to sample the σ^2 parameter from its posterior distribution instead of defining the error as in definition (2.10), which would lead to the following definition:

$$\begin{aligned} R^2 &= \frac{\text{Var}(\sum_{i=1}^n \hat{y}_i^s)}{\text{Var}(\sum_{i=1}^n \hat{y}_i^s) + \sigma_s^2} \\ &= \frac{\boldsymbol{\beta}_s^\top \boldsymbol{\Sigma_{XX}} \boldsymbol{\beta}_s}{\boldsymbol{\beta}_s^\top \boldsymbol{\Sigma_{XX}} \boldsymbol{\beta}_s + \sigma_s^2}, \qquad i = 1, \dots, n, \end{aligned} \tag{2.11}$$

where $\hat{y}_i^s = E(y \mid X_i, \theta^s)$, and $\theta^s, s = 1, \dots, S$ are draws from the posterior parameter distribution.

The predictors in definition (2.10) and definition (2.11) could also be taken as random (Gelman *et al.*, 2017). The predictors are then called stochastic predictors. Using the sample covariance estimate provides then just an estimate of the true covariance structure. With stochastic predictors, there is an additional uncertainty in the R^2 formula that can have a large influence on the R^2 and the LMG values.

In practice, definition (2.11) and definition (2.10) should lead to similar values in the standard linear model. In my opinion, it is more reasonable to take the full Bayesian route by sampling σ^2 of its posterior distribution. This approach provides the opportunity to include prior information about σ^2 directly into to R^2 calculations. The LMG calculations in the examples of this master thesis will therefore be based on definition (2.11). A benefit of definition (2.10) is that it also works for generalized linear models where we often have no separate variance parameter.

The denominator of R^2 is no longer fixed in definition (2.10) and in definition (2.11). We can therefore no longer interpret an increase in R^2 as an improved fit to a fixed target (Gelman *et al.*, 2017). The unfixed denominator seems to be problematic for the LMG formula in the Bayesian framework. However, in the linear model it is possible to calculate the R^2 of all submodels from the parameters of the model inlcuding all predictors (full-model) and the covariance matrix of the predictors. Therefore, all submodels of a posterior sample are compared to the same fixed value. A possible way to get the R^2 of the submodels from the full-model is shown in the next section.

2.3 Conditional variance formula

For two predictors, definition (2.9) simplifies to

$$\text{Var}(y) = \beta_1^2 \text{Var}(X_1) + 2\beta_1\beta_2 \text{Cov}(X_1, X_2) + \beta_2^2 \text{Var}(X_2) + \sigma^2, \tag{2.12}$$

When predictor X_1 is the only one in the model, the explained variance includes the variance of the predictor itself, the whole covariance term, and some of the contribution of the variance of X_2 in equation (2.12) additionaly . In mathematical notation, that is

$$\text{svar}(X_1 \mid \emptyset) = \beta_1^2 \operatorname{Var}(X_1) + 2\beta_1\beta_2 \operatorname{Cov}(X_1, X_2) + \beta_2^2 \operatorname{Var}(X_2)\rho_{12}^2.$$

The contribution of the second regressor is then simply the difference to the total explained variance (Grömping, 2007).

In the general case with p regressors, the conditional variance formula (2.13) can be used to calculate the R^2 of all submodels. For example, the conditional variance formula can be used to specify the conditional distribution of a multivariate normal distribution $\mathbf{Y}$.

The elements of the vector $\mathbf{Y}$ are reordered as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \mathbf{Y}_1 \in \mathbb{R}^q, \mathbf{Y}_2 \in \mathbb{R}^{p-q}.$$

The joint distribution is a multivariate normal distribution with elements

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right), \ \boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^T,$$

the conditional distribution is normally distributed again with mean

$$\mathrm{E}(\mathbf{Y}_1|\mathbf{y}_2) = \boldsymbol{\mu}_1 \ + \ \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Y}_2 \ - \ \boldsymbol{\mu}_2),$$

and the conditional variance is

$$\operatorname{Var}(\mathbf{Y}_1|\mathbf{y}_2) = \boldsymbol{\Sigma}_{11} \ - \ \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{2.13}$$

The aim is to calculate R^2 of a submodel containining the predictors $\mathbf{X}_{q\ldots p}$, and the regression coefficients $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_p)$ without the intercept. The regression coefficients are further separated in $\boldsymbol{\beta}_{1,\ldots,q-1}^\top = (\beta_1, \ldots, \beta_{q-1})$ and $\boldsymbol{\beta}_{q,\ldots,p}^\top = (\beta_q, \ldots, \beta_p)$.

As in the multivariate normal distribution example above, the covariance matrix of p predictors is written as

$$\operatorname{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{XX}} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{p\times p},$$

$$\begin{aligned} \text{where} \quad \boldsymbol{\Sigma}_{11} &= \operatorname{Cov}(\mathbf{X}_{1,\ldots,q-1}, \mathbf{X}_{1,\ldots,q-1}), \\ \boldsymbol{\Sigma}_{12} &= \operatorname{Cov}(\mathbf{X}_{1,\ldots,q-1}, \mathbf{X}_{q,\ldots,p}), \\ \boldsymbol{\Sigma}_{22} &= \operatorname{Cov}(\mathbf{X}_{q,\ldots,p}, \mathbf{X}_{q\ldots p}). \end{aligned}$$

The conditional variance of the predictors $\mathbf{X}_{1,\ldots,q-1}$ given the predictors $\mathbf{X}_{q,\ldots,p}$ is then

$$\operatorname{Cov}(\mathbf{X}_{1,\ldots,q-1} \mid \boldsymbol{x}_{q,\ldots,p}) = \boldsymbol{\Sigma}_{11} \ - \ \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

The total explained variance of the full-model containing $\mathbf{X}_{1\ldots p}$ omits simply the σ^2 parameter in (2.9) , which is

$$\text{evar}(\mathbf{X}_{1,\ldots,p}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}.$$

The explained variance of a submodel can be calculated by subtracting the explained variance of the not-in-the-model-included-predictors that is not explained by in-the-model-included-predictors from the total explained variance. The variance that is not explained by in-the-model-included-predictors is given by the variance of the not-in-the-model-included predictors conditional on the in-the-model-included-predictors. The explained variance of a submodel containing predictors $\mathbf{X}_{q,\dots,p}$ can therefore be written as

$$\text{evar}(\mathbf{X}_{q\dots p}) = \text{evar}(\mathbf{X}_{1,\dots,p}) - \boldsymbol{\beta}_{1,\dots,q-1}^{\top} \operatorname{Cov}(\mathbf{X}_{1,\dots,q-1} \mid \boldsymbol{x}_{q\dots p})\boldsymbol{\beta}_{1,\dots,q-1}. \tag{2.14}$$

To gain the the R^2 value of the submodel, it is necessary to divide the explained variance by the total variance, which is

$$\text{evar}(\mathbf{X}_{q,\dots,p}) / \operatorname{Var}(\mathbf{Y}),$$

where $\operatorname{Var}(\mathbf{Y})$ is definied as $\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{\mathbf{XX}}\boldsymbol{\beta} + \sigma^2$.

A posterior density distribution is obtained for the regression parameters in the Bayesian regression setting. The LMG formula requires calculation of the R^2 values for all $2^p - 1$ submodels. Samples from the joint posterior paramters of the full-model are used to calculate the explained variance of the submodels. For each sample, the conditional variance formula is used to obtain the R^2 of the $2^p - 1$ submodels. The non-negative property and the dependence of the parameters from the submodels to each other is then respected for each sample.

Instead of using the conditional mean formula to get the R^2 of the submodels, it would be possible to fit a separate Bayesian model for each submodel. An R^2 distribution can easily be built for each submodel by using definition (2.10) or definition (2.11). However, the problem is how to calculate the LMG values out of these R^2 distributions. If we just sample independently from the R^2 distributions, the dependence of the paramter values of the submodels to each other is ignored. We would have many possibly true parameter values of a predictor in the same LMG comparison. It would then also be possible that the R^2 decreases when adding predictors. Another drawback is that it would be much more time-consuming to fit a separate Bayesian model for each submodel. Using the conditional variance formula on the full-model allows to calculate LMG values in the Bayesian framework in a reasonable time exposure. Depending on the number of predictors and the number of posterior samples, the calculations still take some time in the Bayesian framework. For stochastic predictors, the computation time is multiplied by the number of covariance samples.

2.4 Bayesian Regression

The following section provides a brief introduction to Bayesian regression. It further shows that assuming stochastic or non-stochastic predictors results in the same posteriors for the regression parameters under some assumptions. It is summarized from the book *Bayesian Analysis for the Social Sciences* (Jackman, 2009).

In regression analysis, we are interested in the dependence of $\boldsymbol{y}$ on $\mathbf{X}$. The conditional mean of a continuous response variable $\boldsymbol{y} = (y_1, \dots, y_n)^{\top}$ is related to a $n \times k$ predictor matrix $\mathbf{X}$ via a linear model,

$$\text{E}(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown regression coefficients.

Under some assumptions about the density, conditional independence and homoskedastic variances, the regression setting can be written as

$$\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

Under the assumption of weak exogeneity and conditional independence, the joint density of the data can be written as

$$p(\boldsymbol{y}, \mathbf{X} \mid \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\theta}_{y|x})\, p(\mathbf{X} \mid \boldsymbol{\theta}_x),$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x)^\top$.

The weak exogeneity assumption implicates that the whole information about $\boldsymbol{y}_i$ is contained in x_i and $\boldsymbol{\theta}_{y|x}$. Knowledge of the parameters $\boldsymbol{\theta}_{x_i}$ provides no additional information about $\boldsymbol{y}_i$. The interest of regression is mostly on the posterior parameters $\boldsymbol{\theta}_{y|x}$. These posterior densities are proportional to the likelihood of the data multiplied by the prior density. The joint density $p(\boldsymbol{y}, \mathbf{X} \mid \boldsymbol{\theta})$ is used to learn about the posterior parameters, via Bayes Rule

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathbf{X}) \propto p(\boldsymbol{y}, \mathbf{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

The dependence of $\boldsymbol{y}$ on $\mathbf{X}$ is captured in the parameters $\boldsymbol{\theta}_{y|x} = (\beta, \sigma^2)$. Under the assumption of independent prior densities about $\boldsymbol{\theta}_{y|x}$ and $\boldsymbol{\theta}_x$ the posterior distribution of the parameters can be written as

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}_x \mid \boldsymbol{y}, \mathbf{X}) = \frac{p(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y} \mid \mathbf{X})} \times \frac{p(\mathbf{X} \mid \boldsymbol{\theta}_x)\, p(\boldsymbol{\theta}_x)}{p(\mathbf{X})}. \tag{2.15}$$

The factorization in equation 2.15 shows, that under the above mentioned assumptions, the posterior inference about the parameters $\boldsymbol{\theta}_{y|x} = (\beta, \sigma^2)$ is independent from the inference about $\boldsymbol{\theta}_x$ given data $\mathbf{X}$. This also means that the assumptions about $\mathbf{X}$ being non-stochastic or stochastic result in the same posterior density of $\boldsymbol{\theta}_{y|x}$. In the case of non-stochastic regressors, $p(\mathbf{X})$ and $\boldsymbol{\theta}_x$ drop out of the calculations. For stochastic predictors, it means, that given $\mathbf{X}$, nothing more can be gained about $\theta_{y|x} = (\boldsymbol{\beta}, \sigma^2)$ from knowing $\boldsymbol{\theta}_x$.

The focus of regression is on $\boldsymbol{\theta}_{y|x} = (\boldsymbol{\beta}, \sigma^2)$, for which it does not matter whether we assume fixed or stochastic predictors under the above mentioned assumptions. The variance of the predictors is also incorporated in the LMG formula. The LMG formula may be especially interesting for continuous predictors, which often are of stochastic nature. Grömping (2006) recommends in most cases to use the non fixed regressor option when calculating bootstrap confidence intervals. Therefore, the information about $\boldsymbol{\theta}_x$ would also be relevant for stochastic regressors. As seen in equation (2.15), inference about $\boldsymbol{\theta}_x$ is independent from inference about $\boldsymbol{\theta}_{y|x}$. If there are stochastic predictors and we use the sample estimate of the covariance matrix, we do not incorporate the uncertainty of the estimate. Because the explained variance is calculated by $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}$, inference about $\boldsymbol{\theta}_x$ seems to be equally important as inference about $\boldsymbol{\theta}_{y|x}$ for stochastic predictors. If the distribution of the $p(\mathbf{X})$ is known, the $\boldsymbol{\theta}_x$ could be estimated. However, the computation time is then much higher, because the whole LMG calculation need to done for each posterior covariance sample of the predictors. Depending on the number of predictors this would be very time-consuming. In most cases, the problem is that the distribution of the $\mathbf{X}$ is unknown. As a practical solution, nonparametric bootstrapping of the covariance matrix could be used to include the uncertainty of the stochastic predictors in the LMG calculations. Again, it would be necessary to do the LMG calculations for each bootstrap sample of the covariance matrix. There exist also different covariance estimators. The shrinkage method may be an interesting estimator with some nice properties (Schäfer and Strimmer, 2005).

Chapter 3

Examples

The following chapter presents the Bayesian LMG implementation by two examples. Simulated data is used for the first example. Empirical data is used for the second example.

3.1 Simulated Data

We assume a simple model for the first example:

$$
\begin{aligned}
&Y_i \sim \mathcal{N}(\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4,\, \sigma^2),\\
&\beta_1 = 0.5, , \beta_2 = 1,\, \beta_3 = 2,\, \beta_4 = 0,\, \sigma^2 = 1\\
&\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4 \sim \mathcal{N}(0, 1)
\end{aligned}
$$

The values of the four predictors are sampled from a standard normal distribution. These values are then multiplied by the regression coefficients. A standard normal distributed error is added to obtain the dependent variable. Fifty observations were sampled. The data generating R-code A.1 can be found in the Appendix.

The model is fitted using the `rstanarm` package (Stan Development Team, 2016) with the default priors for the regression and σ^2 parameters. The exact command can be found in R-code A.2 These default priors are called 'weakly informative priors', because they take into account the order of magnitude of the variables by using the variance of the observed data. Information about these priors can be found in Stan Development Team (2017). A burn-in period of 20000, a sample size of 20000, and a thinning of 20 were chosen, resulting in a posterior sample size of 1000. The first few posterior samples are shown in Table 3.1.

For each posterior sample of the parameters, the R^2 value was calculated. The R^2 of the submodels was then calculated by the conditional variance formula for each posterior sample. The R-code is found in A.3. The resulting R^2 values of the first few posterior samples are shown in Table 3.2. The thinning is reasonable in this case to reduce the computational burden and to still obtain an appropriate posterior of the R^2 values (Link and Eaton, 2012).

The `hier.part` package was used to calculate the LMG value for each posterior sample. The LMG posteriors are shown in Table 3.3 . The independent component (I) represents the LMG value. The joint contribution (J) represents the difference from the independent component to the explained variance of the model containing only the predictor itself (T). Assuming stochastic or non-stochastic regressors has an influence on the uncertainty of the LMG values.

At first, non-stochastic regressors were assumed. The resulting LMG values and joint contributions with a 95% credible interval are shown in Table 3.3. An option to display the resulting LMG distribution is shown in Figure 3.1. Using the default weakly informative priors, the LMG distributions obtained from the Bayesian framework were very similar to the bootstrap confidence intervals, assuming non-stochastic predictors of the LMG estimates obtained from the `relaimpo` package, as shown in Table 3.4.

Table 3.1: Samples from the posterior distributions of the regression parameters

	x1	x2	x3	x4	sigma
sample 1	0.830	0.471	0.689	0.895	1.269
sample 2	1.048	0.998	0.836	0.660	1.074
sample 3	0.986	0.762	0.913	0.703	0.827
sample 4	0.795	0.969	0.789	0.472	1.106
sample 5	1.294	0.943	0.850	0.383	1.037
sample 6	0.817	1.017	0.907	0.825	1.096
sample 7	1.029	1.056	1.073	0.598	0.917
sample 8	1.013	0.851	0.729	0.938	0.946
sample 9	1.215	0.895	0.748	0.825	0.930
sample 10	0.995	1.017	0.832	0.830	0.870

Table 3.2: R^2 for all submodels for the first six posterior samples

	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6
none	0.000	0.000	0.000	0.000	0.000	0.000
x1	0.125	0.173	0.192	0.102	0.315	0.076
x2	0.028	0.133	0.078	0.161	0.084	0.158
x3	0.112	0.111	0.193	0.120	0.099	0.148
x4	0.253	0.177	0.224	0.134	0.075	0.257
x1 x2	0.180	0.373	0.324	0.319	0.473	0.283
x1 x3	0.264	0.317	0.430	0.247	0.455	0.249
x1 x4	0.402	0.374	0.443	0.251	0.411	0.352
x2 x3	0.163	0.293	0.321	0.336	0.220	0.367
x2 x4	0.260	0.266	0.265	0.252	0.136	0.357
x3 x4	0.329	0.259	0.372	0.226	0.156	0.363
x1 x2 x3	0.360	0.596	0.645	0.547	0.682	0.543
x1 x2 x4	0.425	0.519	0.526	0.418	0.535	0.492
x1 x3 x4	0.501	0.482	0.630	0.366	0.527	0.479
x2 x3 x4	0.348	0.388	0.451	0.394	0.250	0.512
all	0.551	0.694	0.780	0.608	0.715	0.692

Table 3.3: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Variable	**I**	**J**	**Total**
x1	0.236 (0.134, 0.35)	-0.064 (-0.076, -0.047)	0.172 (0.083, 0.278)
x2	0.135 (0.056, 0.235)	-0.036 (-0.052, -0.019)	0.098 (0.031, 0.194)
x3	0.195 (0.094, 0.307)	-0.029 (-0.044, -0.015)	0.165 (0.073, 0.277)
x4	0.171 (0.073, 0.285)	0.038 (0.023, 0.051)	0.21 (0.098, 0.33)

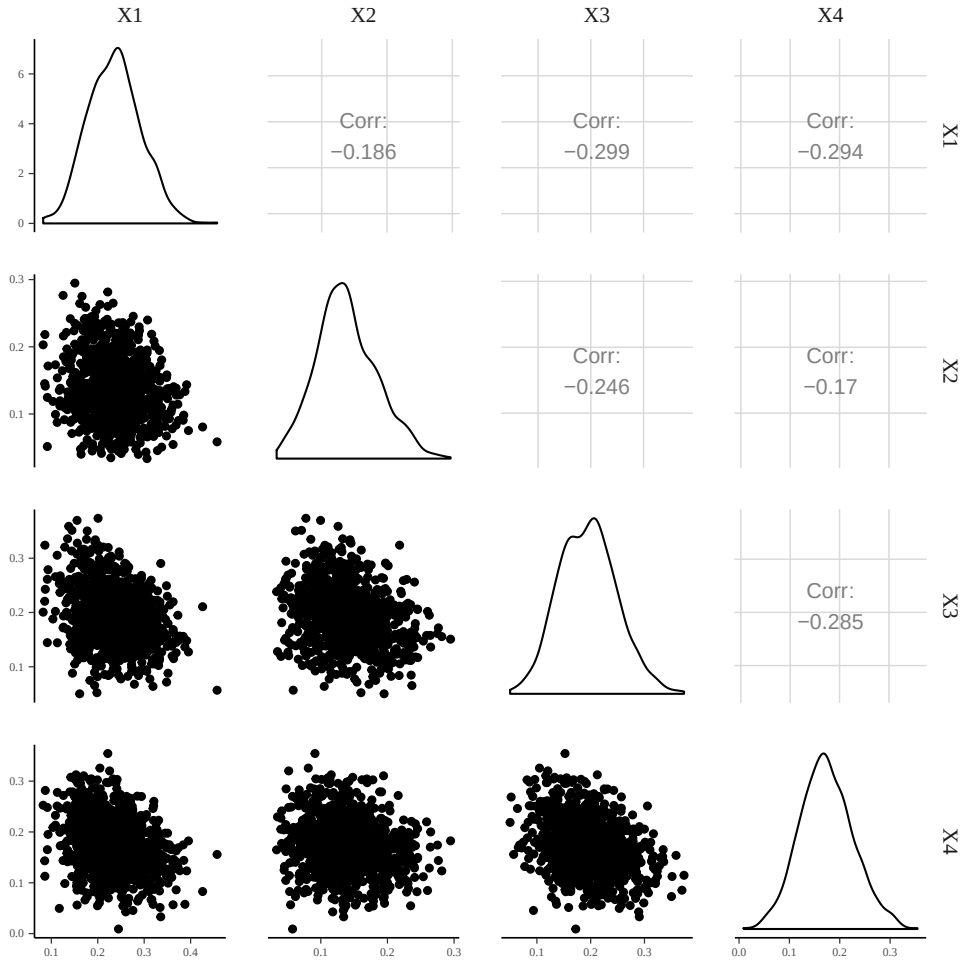

Figure 3.1: Posterior distribution of LMG values.

Table 3.4: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

	LMG value (95%-CI)	
Variable	Relaimpo	**Bayesian framework**
x1	0.244 (0.147, 0.37)	0.236 (0.134, 0.35)
x2	0.145 (0.067, 0.244)	0.135 (0.056, 0.235)
x3	0.198 (0.105, 0.319)	0.195 (0.094, 0.307)
x4	0.18 (0.094, 0.292)	0.171 (0.073, 0.285)

In this example, we know, that the predictor values were sampled from a normal distribution. It would therefore be more reasonable to assume stochastic predictors. Under the assumption of weak exogeinity and conditional independence, the posterior distributions of the regression parameters $\boldsymbol{\beta}$ are valid for non-stochastic and stochastic predictors. However, the uncertainty about the LMG values needs to include the uncertainty about the covariance matrix. If we know the distribution of the predictors we can incorporate this information and obtain the posterior distribution of the covariance matrix. The package `Jags` was used for inference about the covariance matrix in a Bayesian way. As an alternative, non-parametric bootstrap was used for inference about the covariance matrix.

Using the bootstrap samples of the covariance matrix or samples from the posterior covariance matrix produced very similar LMG distributions. Bootstrap seems to be a valuable option for stochastic predictors when the distribution of the predictors is unknown. Even when the distribution is known, the difference seems to be tiny. A benefit of going the full Bayesian way is that we can also include prior knowledge about the covariance matrix. Using the default priors further produced very similar LMG distribution as using the non-parametric bootstrap option of the `relaimpo` package. Table 3.5 shows the LMG values of these approaches. For stochastic predictors, in contrast to non-stochastic predictors, the uncertainty about the covariance matrix is reflected in the larger credible intervals. Even when the exact regression parameters were known, there would a lot of uncertainty in the LMG values caused by the uncertainty about the covariance matrix.

Table 3.5: LMG values assuming stochastic predictors with 95% CI.

	Relaimpo	**Bayesian framework**	
Variable		nonparameteric bootstrap	covariance inference
x1	0.244 (0.092, 0.415)	0.236 (0.115, 0.364)	0.23 (0.116, 0.375)
x2	0.145 (0.053, 0.264)	0.139 (0.057, 0.245)	0.147 (0.059, 0.256)
x3	0.198 (0.082, 0.368)	0.201 (0.098, 0.332)	0.205 (0.08, 0.339)
x4	0.18 (0.073, 0.32)	0.167 (0.058, 0.297)	0.155 (0.046, 0.292)

Table 3.6: Variable description

Variable Name	Description
paragrap	scores on paragraph comprehension test
general	scores on general information test
sentence	scores on sentence completion test
wordc	scores on word classification test
wordm	scores on word meaning test

3.2 Empirical Data

In the following section, the Bayesian LMG implementation is applied on an empirical dataset containing test scores of pupils (N = 301) from a study by Holzinger and Swineford (1939) available in the R package `MBESS` (Kelley, 2017). This dataset was used in Nimon *et al.* (2008) to present commonality analysis, which is another variance decomposition technique. Scores from a paragraph comprehension test (paragrap) were predicted by four verbal tests: general-information (general), sentence-comprehension (sentence) , word-classification (wordc), and word-meaning (wordm) (Table 3.6).

The aim of the regression analysis was to determine the association between verbal ability and paragraph comprehension. An overview of the data is shown in Figure 3.3. The regression results are shown in Table 3.7). A novice researcher may wrongly conclude, that there is little association between the "non-significant" predictors (general information and word-classification) and paragraph comprehension. Given the other predictors are already included in the model, the predictors seem not to provide much information about the expected paragraph comprehension ability. However, it should not be concluded from this regression table, that the association between any of these "non-significant" predictors and the dependent variable is unimportant. As shown in Figure 3.3, the correlations between the predictors are rather high. The LMG metric may therefore provide new information about the importance of each predictor.

The Bayesian regression model was fitted in `rstanarm`. The default priors were used for the regression coefficients and the σ^2 parameter. A burn-in period of 20000, a sample size of 20000, and a thinning of 20 resulted in a posterior sample size of 1000. The first few posterior samples are shown in Table 3.8. The resulting R^2 of these posterior samples are shown in Table 3.9. The LMG values were calculated by using `hier.part`. The independent component (I), joint contribution (J), and total explained variance in a one-predictor model (T) are shown in Table 3.10. Sentence-comprehension and word-meaning seem to be the most important predictors by applying the LMG metric. However, none of the predictors seem to be unimportant. The joint contributions of each predictor were quite large.

For comparison purposes, the LMG metric was additionally calculated with the `relaimpo` package using parametric bootstrapping. The confidence intervals of `relaimpo` were almost identical to the credible intervals of the Bayesian framework (Table 3.11). Assuming stochastic or non-stochastic predictors resulted also in almost identical uncertainty estimates with such a large sample size (Table3.12).

Table 3.7: Regression of paragraph comprehension on verbal tests.

	Coefficient	95%-confidence interval	*p*-value
Intercept	0.071	from -1.17 to 1.31	0.91
general	0.03	from -0.00 to 0.06	0.084
sentence	0.26	from 0.18 to 0.34	< 0.0001
wordc	0.047	from -0.01 to 0.11	0.14
wordm	0.14	from 0.08 to 0.19	< 0.0001

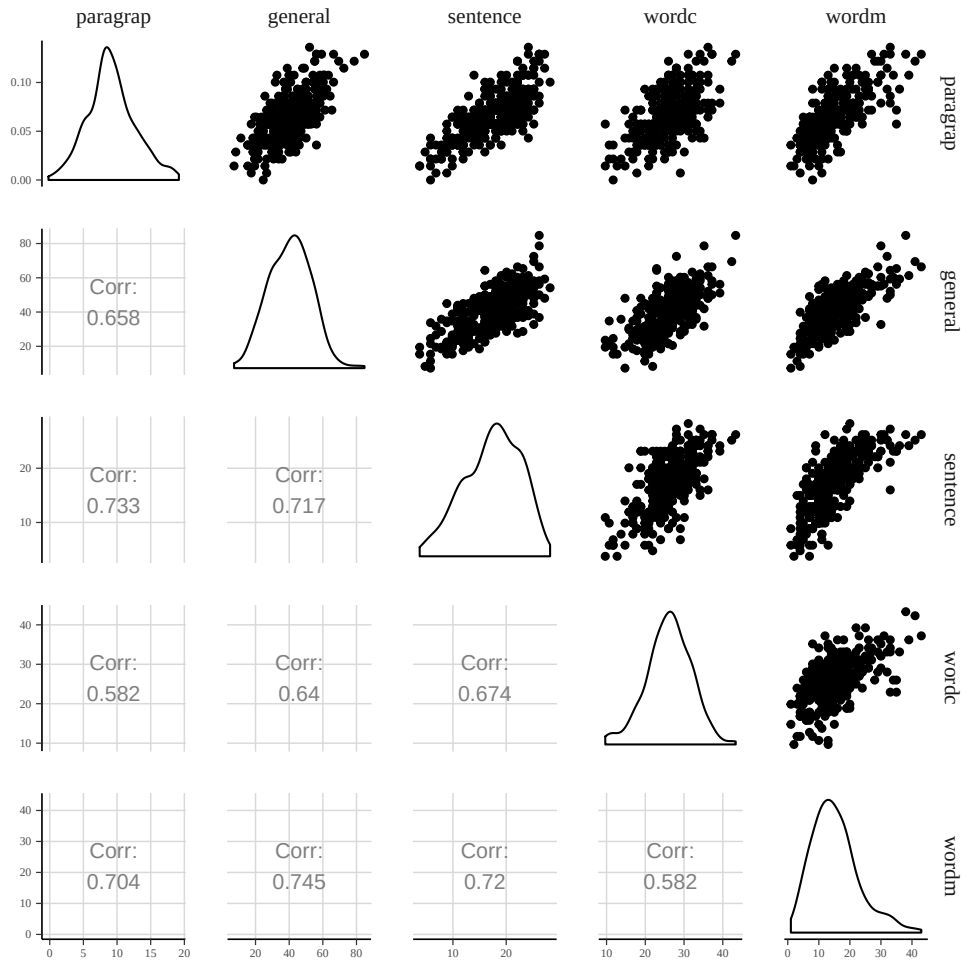

Figure 3.2: Test scores from Holzinger and Swineford's (1939) Study. N=301

Table 3.8: Samples from the posterior distributions of the regression parameters

	general	sentence	wordc	wordm	sigma
sample 1	0.038	0.218	0.043	0.141	2.08
sample 2	0.016	0.356	0.019	0.120	2.26
sample 3	0.037	0.262	0.071	0.122	2.16
sample 4	0.041	0.252	0.026	0.123	2.21
sample 5	0.034	0.287	0.045	0.131	2.21
sample 6	0.041	0.192	0.064	0.136	2.17
sample 7	0.009	0.330	0.047	0.102	2.27
sample 8	0.028	0.258	0.017	0.136	2.36
sample 9	0.029	0.262	0.037	0.136	2.23
sample 10	0.024	0.289	0.037	0.167	2.33

Table 3.9: R^2 for all submodels for the first six posterior samples

	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6
none	0.000	0.000	0.000	0.000	0.000	0.000
general	0.449	0.395	0.453	0.428	0.443	0.437
sentence	0.519	0.565	0.551	0.511	0.554	0.487
wordc	0.335	0.309	0.370	0.308	0.343	0.339
wordm	0.510	0.461	0.493	0.470	0.496	0.484
general sentence	0.568	0.581	0.592	0.552	0.590	0.540
general wordc	0.488	0.435	0.507	0.460	0.486	0.480
general wordm	0.553	0.495	0.544	0.516	0.540	0.530
sentence wordc	0.535	0.569	0.572	0.521	0.567	0.509
sentence wordm	0.598	0.605	0.609	0.572	0.613	0.565
wordc wordm	0.550	0.500	0.554	0.507	0.543	0.532
general sentence wordc	0.572	0.582	0.600	0.554	0.593	0.547
general sentence wordm	0.608	0.606	0.619	0.581	0.620	0.577
general wordc wordm	0.570	0.514	0.574	0.531	0.562	0.552
sentence wordc wordm	0.604	0.605	0.620	0.574	0.618	0.575
all	0.611	0.606	0.626	0.582	0.623	0.583

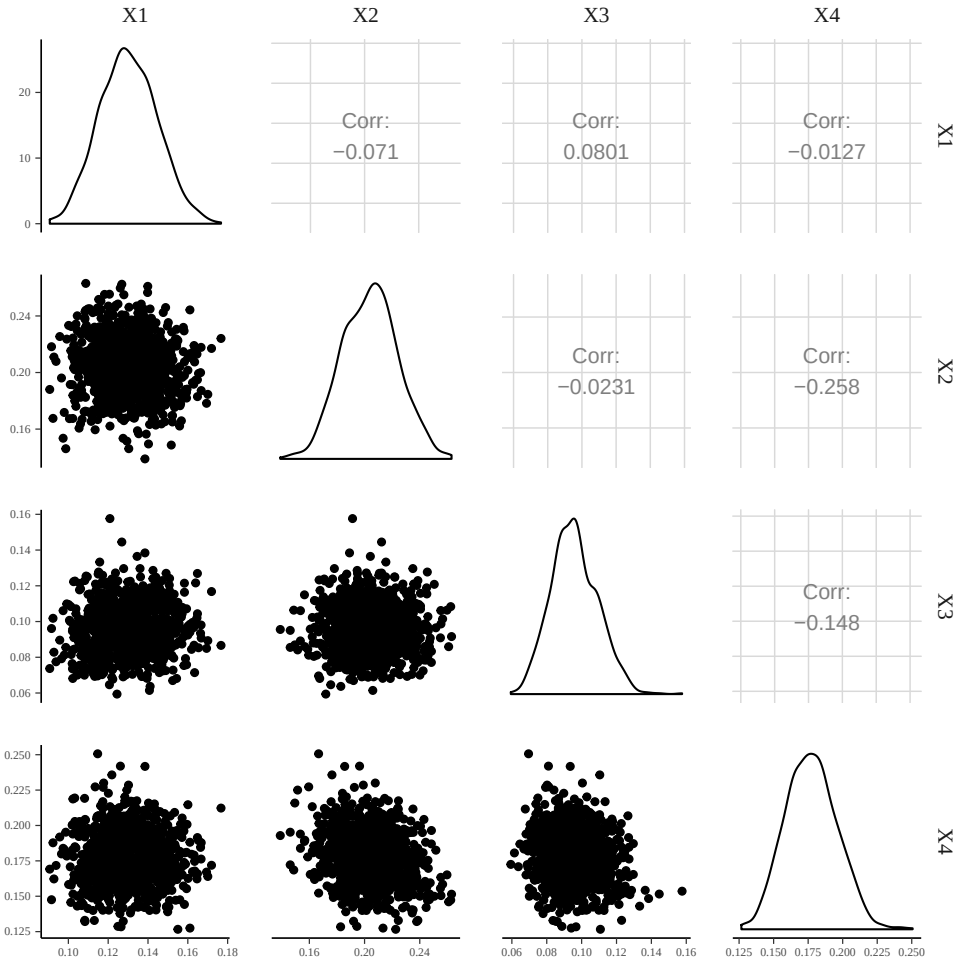

Figure 3.3: LMG posterior distribution of different verbal ability tests

Table 3.10: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Variable	**I**	**J**	**Total**
general	0.13 (0.103, 0.159)	0.298 (0.256, 0.332)	0.428 (0.36, 0.489)
sentence	0.205 (0.167, 0.244)	0.327 (0.292, 0.358)	0.532 (0.466, 0.59)
wordc	0.095 (0.071, 0.123)	0.238 (0.196, 0.274)	0.334 (0.267, 0.396)
wordm	0.177 (0.143, 0.213)	0.314 (0.279, 0.345)	0.491 (0.43, 0.551)

Table 3.11: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

	LMG value (95%-CI)	
Variable	Relaimpo	**Bayesian framework**
general	0.131 (0.104, 0.161)	0.13 (0.103, 0.159)
sentence	0.206 (0.167, 0.25)	0.205 (0.167, 0.244)
wordc	0.096 (0.074, 0.129)	0.095 (0.071, 0.123)
wordm	0.178 (0.144, 0.215)	0.177 (0.143, 0.213)

Table 3.12: Variance decomposition for Stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

	LMG value (95%-CI)	
Variable	Relaimpo	**Bayesian framework**
general	0.131 (0.105, 0.164)	0.13 (0.097, 0.163)
sentence	0.206 (0.168, 0.25)	0.205 (0.167, 0.244)
wordc	0.096 (0.074, 0.124)	0.097 (0.071, 0.125)
wordm	0.178 (0.144, 0.222)	0.176 (0.142, 0.214)

Chapter 4

Extension to longitudinal data

In the following chapter some extensions of the LMG formula beyond the simple linear regression model are shown. The focus is on repeated measurements models. These models extend the simple linear regression by allowing intra-subject correlation between repeated measures.

The dependence of within-subject measurements can be modeled by including random effects (mixed model) or by assuming correlated errors within a subject (marginal model). A mixed model can be extended by including a random slope per subject, allowing for less restrictive longitudinal shapes. The marginal approach can get more freedom by different specified covariance matrices of the error terms. An unstructured covariance matrix, where no restriction are imposed, allows for the most freedom. However, depending on the number of repeated measurements and the sample size the covariance matrix can get too large to make reasonable inference about it (Fitzmaurice *et al.*, 2011).

The extension of the LMG formula in the Bayesian framework to longitudinal models is restricted to models where the conditional variance formula can be applied to get the explained variance of the submodel from the regression parameters of the full model. The focus is therefore on the fixed predictors and not on the random effects. The conditional variance formula can be used in the marginal models, where only the fixed effects are modeled anyway. In the mixed model framework, the conditional variance formula is applicable to random intercepts models. For random-slope models there are at least some difficulties involved, if it is possible at all, to get the explained variance of the submodel. This chapter shows the Bayesian LMG Implementation on a random intercept model and on a repeated measurement model with an unstructured covariance matrix.

4.1 Random intercept model

The first example concerns a simple random intercept model with time-varying predictors. Different R^2 metrics exist for linear mixed models. The variance of a random intercept model with regression parameter $\boldsymbol{\beta}$ can be written as

$$\mathrm{Var}(y) = \sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2, \tag{4.1}$$

where $\sigma_f^2 = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}$, σ_α^2 is the variance of the random intercept and σ_ϵ^2 represents the error variance (Nakagawa and Schielzeth, 2013).

An R^2 that is guaranteed to be positive is defined in Nakagawa and Schielzeth (2013) as

$$R^2_{\mathrm{LMM}} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}. \tag{4.2}$$

It is theoretically possible that the R^2_{LMM} decreases when adding predictors (Nakagawa and Schielzeth, 2013). By adding predictors σ^2_f should always increase and σ^2_ϵ decrease. However, the σ^2_α may also increase a little bit and the total R^2 may then be a little bit lower. The R^2 can not decrease by using the conditional variance formula on the full model to calculate the R^2 of the submodels, because the total variance is fixed. The results should be the same, as if we would fit a new model by maximum likelihood for each submodel and compare the explained variance of the fixed effects to the explained variance of the full model. In the Bayesian framework, the conditional variance formula is needed to account for the interdependence of the submodels to each other. The total variance of the full model can be calculated as $\text{Var}(y) = \text{Var}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}b)+\sigma^2$ or by using samples of σ^2_α as in (4.1). The error term could again be sampled or calculated as in (**??**). In the following examples, (4.1) is used and σ^2_α and σ^2_ϵ are sampled from their posterior distribution.

In repeated measurement studies, the focus is often in within-subject changes. The between-subject variance, estimated with the random intercept term, is of minor importance. The more important question may be, how much variance the fixed predictors explain, compared to the within subject error, which is

$$R^2_{\text{repeated}} = \frac{\sigma^2_f}{\sigma^2_f + \sigma^2_\epsilon}, \tag{4.3}$$

The square root of this term is known under the name correlation within subjects in Bland and Altman (1995). Often, there are between-subject and within-subject predictors in a model. If we are interested in the within-subject effect, we can use a model including only the between-subject predictors as the null-model.

The following example shows a simple random intercept model with time-varying predictors. The main question is which within-subject predictors are the most important ones. The between-subject variance is of minor importance.

The data are simulated from the following regression setting with $m = 4$ timepoints and $n = 20$ number of subjects ,

$$Y_{i,j} \sim \mathcal{N}(\beta_0 + x_{1_{i,j}}\beta_1 + x_{2_{i,j}}\beta_2 + x_{3_{i,j}}\beta_3 + x_{4_{i,j}}\beta_4 + \alpha_i,\, \sigma^2), \qquad \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, m \end{matrix}$$

where $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 2$ $\beta_4 = 2$, $\sigma^2 = 1$, $\alpha_i \sim \mathcal{N}(0, \sigma^2_\alpha)$, $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

The random intercept effect is of minor interest. The R^2 of the models is calculated according to the formula of repeated measure correlation (4.3). Most of the within-subject variance is explained by the predictors (Table 4.1). The credible intervals are very narrow. For information about the between-subject variance term, we can look at the posterior distribution of the random intercept variance term.

In the second part, the random intercept is directly included in the total variance calculation of the R^2 values. There is a large between-subject variance in this simulated dataset (Table 4.2). The LMG values including the between subject variance are therefore much lower. The credible intervals are as well much wider, because the uncertainty about the between-subject variance is included.

In my opinion we can get more useful information from separating the between-subject and within-subject variance components in this simple case. Note that we assumed non stochastic predictors. Otherwise, the credible intervals would be larger. In general, it seems more reasonable to assume stochastic time-varying predictors. The variance could then be estimated by non-parametric bootstrap, resampling whole subjects (all repeated measurements of a subject).

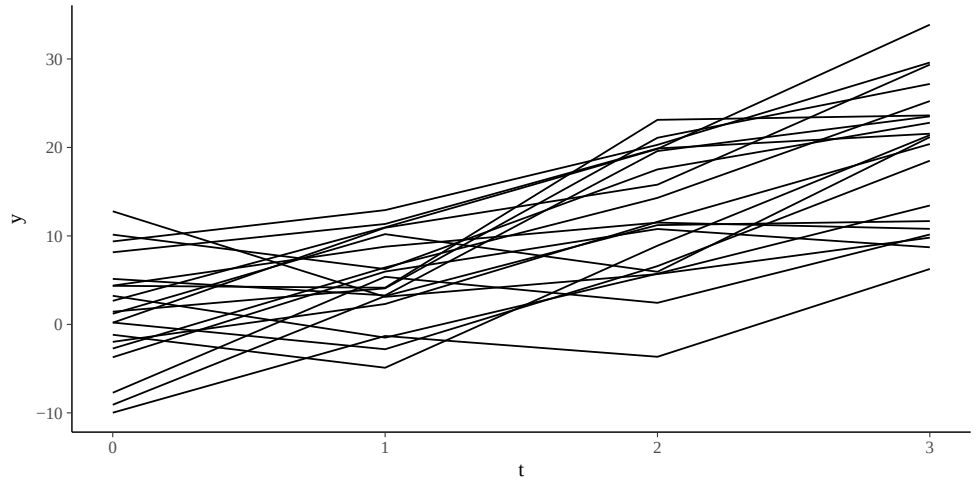

Figure 4.1: Individual trajectories of simulated random intercept model

Table 4.1: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Variable	I	J	Total
x1	0.204 (0.202, 0.205)	0.507 (0.505, 0.509)	0.711 (0.708, 0.714)
x2	0.19 (0.189, 0.192)	0.473 (0.472, 0.475)	0.663 (0.66, 0.667)
x3	0.304 (0.302, 0.305)	0.636 (0.635, 0.636)	0.939 (0.938, 0.941)
x4	0.302 (0.301, 0.304)	0.633 (0.632, 0.633)	0.935 (0.933, 0.937)

4.2 Marginal model

The next example concerns a repeated measurement model with time-varying predictors and an unstructured error covariance matrix. The data are generated from the following model:

$$Y_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, n \tag{4.4}$$

where $\boldsymbol{\Sigma}$ represents an unstructured error covariance matrix, $\mathbf{X}_i$ represents the predictor matrix of size $m \times p$ of subject i.

In the variance calculation we need to take into account that we do not have just one σ^2 parameter, but a covariance matrix $\boldsymbol{\Sigma}$. The diagonal elements of $\boldsymbol{\Sigma}$ represent the variance of each timepoint. The sum of the diagonal elements of $\boldsymbol{\Sigma}$ represents the variance for a whole subject. We can take the mean of diag($\boldsymbol{\Sigma}$) to make the formula compatible with the $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}$ of (**??**), resulting in the total variance term

$$\text{Var}(\mathbf{Y}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta} + \text{mean}(\text{diag}(\boldsymbol{\Sigma})), \tag{4.5}$$

The individual trajectories are shown in Figure 4.2. The resulting LMG values of the predictors are shown in Table 4.3.

Table 4.2: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Variable	I	J	Total
x1	0.077 (0.027, 0.141)	0.198 (0.073, 0.327)	0.275 (0.1, 0.466)
x2	0.084 (0.041, 0.127)	0.195 (0.09, 0.303)	0.279 (0.132, 0.43)
x3	0.118 (0.062, 0.175)	0.253 (0.122, 0.387)	0.372 (0.185, 0.561)
x4	0.123 (0.066, 0.181)	0.253 (0.123, 0.386)	0.376 (0.192, 0.564)

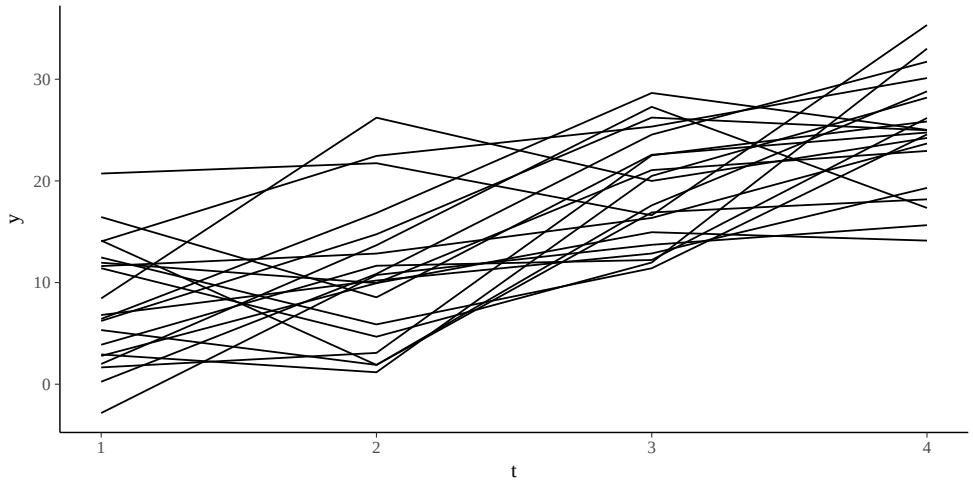

Figure 4.2: Individual trajectories of simulated data with unstructured error covariance matrix

Table 4.3: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Variable	I	J	Total
x1	0.168 (0.139, 0.189)	0.388 (0.321, 0.432)	0.556 (0.461, 0.62)
x2	0.137 (0.111, 0.156)	0.366 (0.302, 0.408)	0.503 (0.413, 0.563)
x3	0.198 (0.162, 0.222)	0.449 (0.369, 0.499)	0.648 (0.532, 0.721)
x4	0.209 (0.17, 0.235)	0.462 (0.379, 0.514)	0.671 (0.548, 0.747)

Chapter 5

Discussion and Conclusion

5.1 Other variable importance metrics in the Bayesian framework

Different variable importance metrics exist (Grömping, 2015). The focus of this master thesis was on the LMG formula. A lot of variable importance metrics are based on the R^2 of the full model compared to the submodels. Instead of the LMG formula, another variable importance metric could as well be used on the R^2 of all submodels for each posterior sample. Commonality analysis (Nimon *et al.*, 2008) and dominance analysis (Grömping, 2015) seem to be interesting variance decomposition metrics. Both provide some extensions to the LMG framework. Different options exist further for the description of the LMG distribution. The `relaimpo` package provides for example some more bootstrap options, like pairwise differences, that could easily be transferred to the Bayesian framework.

5.2 Conclusion

The Bayesian frameworks provides the option to include the uncertainty about parameters. Using the conditional variance formula allows to calculate the R^2 of all the submodels from the posterior parameter distributions of the full model. Instead of fitting $2^p - 1$ models, only the full model needs to be fitted. The interdependence of the submodels to each other is then automatically respected. The R^2 of the submodels does not decrease when adding predictors. The important property of non-negativity shares is then respected in the Bayesian framework.

A disadvantage about calculating the R^2 of all the submodels with the conditional variance formula seems to be the restriction to the linear model. Although this may be a topic of further research. A further disadvantage of the Bayesian framework compared to the classical LMG implementation, are the higher computational costs. For non-stochastic predictors, the calculations are possible in a reasonable amount of time. Parallel computing could be used to speed up the calculations. For stochastic predictors, the computation times will be much higher than in the classical framework.

Assuming non-stochastic or stochastic predictors can have a big impact in small samples on the the uncertainty of the explained variance and the LMG values. Although the posterior regression parameter distributions is the same in both cases (under some assumptions described in chapter 2), the explained variance of a model is directly dependent on the covariance matrix. Inference about the covariance of the predictors $\mathbf{X}$ is therefore an important part when stochastic predictors are assumed. However, this does in general not seem to be an easy problem. Non-parametric bootstrap provides a practical solution in the Bayesian framework.

When the sample size is large enough the classical and the Bayesian framework should lead to very similar values. The Bayesian framework allows to include prior information, that may especially be relevant for small sample sizes. The credible interval of the Bayesian framework,

compared to the confidence intervals, may further be easier to interpret in the mathematically correct way.

A lot of studies are concerned with within-subject changes. The extension of the LMG formula to those kind of problems is not straightforward. It is depending on the complexity of the data. However, for the simple random intercept model the extension seems to be easily possible when the focus is on the fixed effects. In the Bayesian and maximum likelihood framework it seems reasonable to compare only the explained variance of the fixed effects of the submodels against the total variance, containing $\sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2$. Otherwise, there may be problems with the non-negativity property of the shares.

Appendix A

Some additional notes

The eror term in the by Gelman *et al.* (2017) proposed definition of the R^2 is defined as $\text{Var}(\sum_{i=1}^{n} e_n^s)$. I think we could also use $\sum(y - \hat{y}^s)^2/(n-1)$ as an estimate for the error. For the maximum likelihood estimate $\text{Var}(y_i - \hat{y}_i) = \sum(y_i - \hat{y}_i)^2/(n-1)$. This is because the mean of the residuals is 0. When samples of the posterior parameters are used, the mean of the residuals is not excatly zero. $\text{Var}(y_i - \hat{y}_i) = \sum(y_i - \hat{y}_i)^2/(n-1)$ is than a little bit bigger than $\text{Var}(y_i - \hat{y}_i)$. In practice the values should only differ by a very small amount. We do not expect the errors to have a systematic bias. However, the residuals are just a sample of the error. The mean of the residuals must not be excatly 0 when the samples of the posteriors are used for the regression coefficients.

A.1 Code used in chapter 3

The data of the first example of chapter 3 were generated with R-Code A.1

R-Code A.1: Data generation of example 1 chapter 3

```
x1 <- rnorm(50, 0, 1)
x2 <- rnorm(50, 0, 1)
x3 <- rnorm(50, 0, 1)
x4 <- rnorm(50, 0, 1)
# b1 <- 0.5; b2 <- 1; b3 <- 2; b4 <- 0.8
b1 <- 1
b2 <- 1
b3 <- 1
b4 <- 1

y <- b1 * x1 + x2 * b2 + b3 * x3 + b4 * x4 + rnorm(50, 0, 1)

df <- data.frame(y = y, x1 = x1, x2 = x2, x3 = x3, x4 = x4)
```

R-Code A.2: Bayesian regression model for example 1

```
post2 <- stan_glm(y ~ 1 + x1 + x2 + x3 + x4, data = df, chains = 1, cores = 1,
    iter = 40000, thin = 20)
```

R-Code A.3: Function to get R^2 from posterior samples of the full model

```
rtwos <- function(X, post.sample) {
    # X: Predictor data as data frame post.sample: posterior sample as matrix
    # (M[sample_i,]), last position should be sigma paramater.
    X <- cov(X)  #covariance matrix
    # Prepare data frame and rownames by using the combn() function
    lst <- list()
    pcan <- dim(X)[2]
    n <- (2^pcan) - 1
    for (i in 1:pcan) {
        lst[[i]] <- combn(pcan, i)
    }
    var.names <- character(length = 0)
    v <- rownames(X)
    for (i in 1:length(lst)) {
        for (j in 1:ncol(lst[[i]])) {
            cur <- lst[[i]][, j]
            name <- paste0(v[-cur])
            name <- paste(name, collapse = " ")
            var.names <- c(var.names, name)
        }
    }
    var.names <- c(rev(var.names), "all")
    var.names[1] <- "none"
    size <- nrow(post.sample)  # how many samples
```

```
    sig.posi <- ncol(post.sample)
    df.Rtwos <- data.frame(matrix(0, n + 1, size))
    row.names(df.Rtwos) <- var.names
    # fill in R^2 matrix, use posterior samples and calculate submodels
    # according to the conditional variance formula.
    v <- 1:dim(X)[2]
    for (s in 1:size) {
        sample.s <- post.sample[s, -sig.posi]
        Vtot <- sample.s %*% X %*% sample.s  #total explained variance
        count = n
        for (i in 1:(length(lst) - 1)) {
            for (j in 1:ncol(lst[[i]])) {
                cur <- lst[[i]][, j]
                set <- v[-cur]
                matr <- X[cur, cur] - X[cur, set] %*% solve(X[set, set]) %*%
                  X[set, cur]  #conditional variance formula
                var.explain <- sample.s[cur] %*% matr %*% sample.s[cur]
                df.Rtwos[count, s] <- Vtot - var.explain
                count = count - 1
            }
        }
        df.Rtwos[n + 1, s] <- Vtot
        df.Rtwos[, s] <- df.Rtwos[, s]/c(Vtot + post.sample[s, sig.posi]^2)
    }
    return(df.Rtwos)
}
```

R-Code A.4: Bayesian regression model for example 1

```
# helper function to calculate covariance bootstrap sample

bootcov <- function(df, boot.n) {
    len <- nrow(df)
    cov.m <- cov(df)
    l <- dim(cov.m)[1]
    M.boot <- array(NA, c(l, l, boot.n))
    M.boot[, , 1] <- cov(df)
    for (i in 2:boot.n) {
        dfs <- df[sample(1:len, replace = T), ]
        M.boot[, , i] <- cov(dfs)
    }

    return(M.boot)
}


# Function that includes uncertainty about stochastic predictors by
# bootstrapping using bootcov boot.n is the number of bootstrap samples
# takes boot.n times longer to calculate.
```

```
rtwos.boot <- function(df, post.sample, boot.n) {
    # df: Predictor data as data frame post.sample: posterior sample as matrix
    # (M[sample_i,]), last position should be sigma paramater.

    X <- cov(df)  #covariance matrix

    # Prepare data frame and rownames by using the combn() function

    lst <- list()

    pcan <- dim(X)[2]
    n <- (2^pcan) - 1

    for (i in 1:pcan) {
        lst[[i]] <- combn(pcan, i)
    }

    var.names <- character(length = 0)

    v <- rownames(X)

    for (i in 1:length(lst)) {

        for (j in 1:ncol(lst[[i]])) {
            cur <- lst[[i]][, j]
            name <- paste0(v[-cur])
            name <- paste(name, collapse = " ")
            var.names <- c(var.names, name)
        }
    }

    var.names <- c(rev(var.names), "all")

    var.names[1] <- "none"

    size <- nrow(post.sample)  # how many samples

    sig.posi <- ncol(post.sample)

    df.Rtwos <- array(0, c(n + 1, size, boot.n))

    dimnames(df.Rtwos)[[1]] <- var.names

    boot.M <- bootcov(df, boot.n)

    v <- 1:dim(X)[2]

    for (b in 1:boot.n) {
```

```
        X <- boot.M[, , b]

        ########### fill in R^2 matrix, use posterior samples and calculate submodels
        ########### according to the conditional variance formula.

        for (s in 1:size) {

            sample.s <- post.sample[s, -sig.posi]

            Vtot <- sample.s %*% X %*% sample.s  #total explained variance

            count = n

            for (i in 1:(length(lst) - 1)) {

                for (j in 1:ncol(lst[[i]])) {
                  cur <- lst[[i]][, j]
                  set <- v[-cur]
                  matr <- X[cur, cur] - X[cur, set] %*% solve(X[set, set]) %*%
                    X[set, cur]  #conditional variance
                  var.explain <- sample.s[cur] %*% matr %*% sample.s[cur]
                  df.Rtwos[count, s, b] <- Vtot - var.explain
                  count = count - 1
                }
            }

            df.Rtwos[n + 1, s, b] <- Vtot

            df.Rtwos[, s, b] <- df.Rtwos[, s, b]/c(Vtot + post.sample[s, sig.posi]^2)
        }

    }
    return(df.Rtwos)
}
```

R-Code A.5: Bayesian regression model for example 1

```
# Function to include samples of the predictor covariance matrix boot.n =
# number of samples covm = posterior sample of covariance matrix

rtwos.covm <- function(df, post.sample, covm, boot.n) {
    # df: Predictor data as data frame post.sample: posterior sample as matrix
    # (M[sample_i,]), last position should be sigma paramater.

    X <- cov(df)  #covariance matrix

    # Prepare data frame and rownames by using the combn() function

    lst <- list()
```

```
pcan <- dim(X)[2]
n <- (2^pcan) - 1

for (i in 1:pcan) {
    lst[[i]] <- combn(pcan, i)
}

var.names <- character(length = 0)

v <- rownames(X)

for (i in 1:length(lst)) {

    for (j in 1:ncol(lst[[i]])) {
        cur <- lst[[i]][, j]
        name <- paste0(v[-cur])
        name <- paste(name, collapse = " ")
        var.names <- c(var.names, name)
    }
}

var.names <- c(rev(var.names), "all")

var.names[1] <- "none"

size <- nrow(post.sample)  # how many samples

sig.posi <- ncol(post.sample)

df.Rtwos <- array(0, c(n + 1, size, boot.n))

dimnames(df.Rtwos)[[1]] <- var.names

v <- 1:dim(X)[2]


for (b in 1:boot.n) {


    X <- covm[, , b]

    ########### fill in R^2 matrix, use posterior samples and calculate submodels
    ########### according to the conditional expectation formula.

    for (s in 1:size) {

        sample.s <- post.sample[s, -sig.posi]

        Vtot <- sample.s %*% X %*% sample.s  #total explained variance
```

```
            count = n

            for (i in 1:(length(lst) - 1)) {

                for (j in 1:ncol(lst[[i]])) {
                  cur <- lst[[i]][, j]
                  set <- v[-cur]
                  matr <- X[cur, cur] - X[cur, set] %*% solve(X[set, set]) %*%
                    X[set, cur]  #conditional variance
                  var.explain <- sample.s[cur] %*% matr %*% sample.s[cur]
                  df.Rtwos[count, s, b] <- Vtot - var.explain
                  count = count - 1
                }
            }

            df.Rtwos[n + 1, s, b] <- Vtot

            df.Rtwos[, s, b] <- df.Rtwos[, s, b]/c(Vtot + post.sample[s, sig.posi]^2)
        }

    }
    return(df.Rtwos)
}
```

Bibliography

Alexander, D. L., Tropsha, A., and Winkler, D. A. (2015). Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling*, **55**, 1316–1322. 4

Bland, J. M. and Altman, D. G. (1995). Calculating correlation coefficients with repeated observations: Part 1–Correlation within subjects. *BMJ (Clinical research ed.)*, **310**, 446. 18

Chevan, A. and Sutherland, M. (1991). Hierarchical partitioning. *American Statistician*, **45**, 90–96. 4

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied longitudinal analysis*. Wiley. 17

Gelman, A., Goodrich, B., Gabry, J., and Ali, I. (2017). R-squared for Bayesian regression models *. Technical report. 5, 23

Grömping, U. (2006). Relative Importance for Linear Regression in R : The Package relaimpo. *Journal of Statistical Software*, **17**, 1–27. 4, 8

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *American Statistician*, **61**, 139–147. 1, 6

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**, 137–152. 1, 3, 4, 21

Holzinger, K. J. and Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary Educational Monographs*, **48**, 1–91. 13

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley. 7

Kelley, K. (2017). Mbess (version 4.0.0 and higher) [computer software and manual]. R package version 4.0.0 and higher. 13

Kvalseth, T. O. (1985). Cautionary Note about R 2. *The American Statistician*, **39**, 279. 4

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, **3**, 112–115. 9

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, 133–142. 17, 18

Nimon, K., Lewis, M., Kane, R., and Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. *Behavior Research Methods*, **40**, 457–466. 13, 21

Nimon, K. F. and Oswald, F. L. (2013). Understanding the Results of Multiple Linear Regression: Beyond Standardized Regression Coefficients. *Organizational Research Methods*, **00**, 1–25. 1

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, 1–30. 8

Stan Development Team (2016). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.13.1. 9

Stan Development Team (2017). *Stan Modeling Language: User's Guide and Reference Manual*. 9

Walsh, C. and Nally, R. M. (2015). Title Hierarchical Partitioning. Technical report. 4