

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

p-values:
their use, abuse and proper use
illustrated with seven facets

Mäxli Musterli

Version July 27, 2018

Contents

Preface	iii
1 Introduction	1
2 Data	3
2.1 Bayesian Regression	8
3 Examples	11
3.1 Simulated Data	11
3.2 Real World Data	26
4 Discussion and Outlook - Some extenstions	35
4.1 random intercept model	35
4.2 marginal model	39
5 Other variable importance metrics in the Bayesian framework	45
5.1 Conclusion	45
6 Appendix	47
Bibliography	49

Preface

Howdy!

Max Muster
June 2018

Chapter 1

Introduction

The aim 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.

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 Y and a set of explanatory variables. The explanatory variables are also called predictors or covariates. Regression parameters provide us the information how much the response variable is expected to change when a predictor changes by one unit, given all other predictors in the model stay the same. The last subsentence is very important for the correct interpretation of the regression parameters. It shows also 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 clear that it is not an easy task to find the most important predictors in a model. The first question then is: What do we mean by the importance of a predictor? A question that is not easy answered and depending on the research question. [Grömping \(2015\)](#) concludes that there may never be a unique accepted definition of what variable importance is. Different metrics exist to quantify the importance of predictors. These metrics focus on different aspects and with correlated predictors they lead to different conclusions. A summary of the 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 regression models who wish to find an association between predictors and the response variable. In the former case, the aim is only to reduce the error between the predicted values and the real observed values. It does not really matter how we get there. In the other case, we are interested in 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 from the regression output unimportant predictor may be the main cause of the other two predictor values. It therefore may somehow be the most important predictor in this model ([Grömping, 2007](#)).

The causal relationship between the variables is missing in the regression model. Regressing conditional on other variables or using univariate regression models only provide us some parts of the bigger picture about the predictor in a model. Some authors recommend that the variable importance metric is based on both components. Which variable importance metrics are the most useful ones is still an open debate. A convincing theoretical basis is still lacking for all of them. [Grömping \(2015\)](#) recommends to use the existing best practices, until a more profound solution is found. For variance (or generelly goodness of fit) decomposition based importance, she recommends to use LMG enhanced with joint contributions or dominance analysis ([Grömping, 2007](#)).

Chapter 2

Data

Maybe it is the methods section. Here however, we give a couple hints. Note that you can wisely use `preamble`-chunks. Minimal, is likely:

```
library(knitr)
opts_chunk$set(
  fig.path='figure/ch02_fig',
  self.contained=FALSE,
  cache=TRUE
)
```

Defining figure options is very helpful:

```
library(knitr)
opts_chunk$set(fig.path='figure/ch02_fig',
              echo=TRUE, message=FALSE,
              fig.width=8, fig.height=2.5,
              out.width='\\textwidth-3cm',
              message=FALSE, fig.align='center',
              background="gray98", tidy=FALSE, #tidy.opts=list(width.cutoff=60),
              cache=TRUE
)
options(width=74)
```

Notice how in Figure 2.1 everything is properly scaled.

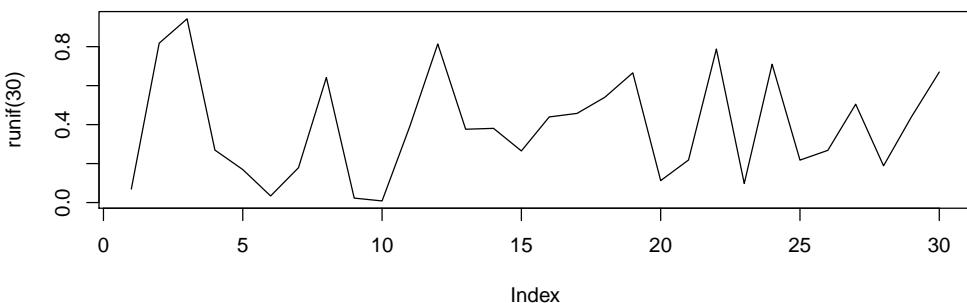


Figure 2.1: Test figure to illustrate figure options used by knitr.

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. Many authors call this a desirable property of a variable importance metric. Marginal and conditional information are incorporated (Grömping, 2015) . The following formulas are taken from Grömping (2015). The same mathematical notations are used.

The following notations for the explained variance (1) and sequentially added variance (2) simplify the notation of the LMG formula.

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

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

, 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 $\text{evar}(S)/\text{Var}(Y)$.

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

$$= \frac{1}{p} \sum_{i=0}^{p-1} \frac{\sum_{\substack{S \subseteq \{2, \dots, p\} \\ n(S)=1}} \text{svar}(\{1\} \mid S)}{\binom{p-1}{i}} \quad (2.4)$$

$$(2.5)$$

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

The different writings of the formulas help to better grasp what is calculated in the LMG metric. The R^2 of the model including all predictors is decomposed. In the top formula 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 middle formula shows that the calculation can be done computationally more efficient. 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 bottom formula 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 .

Chevan and Sutherland 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 predictors itself is included in the model, compared to its independent component (I) is named the joint contribution (J). When R^2 is chosen as the measure of fit the same LMG values as in relaimpo was calculated. However, the hier.part package is only guaranteed to work for 9 predictors.

For the linear model the R^2 is the most widely used goodness-of-fit metric. Different formulas for R^2 exist [Kvalseth \(1985\)](#), all leading to the same value when and intercept is included and the model is fitted by maximum likelihood.

Two popular definitions are:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.6)$$

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

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

When other estimation methods than maximum likelihood are used (2.6) can be negative and (2.7) 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 does not make sense. 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. Because of R^2 bigger than one or the possibility of negative R^2 values [Gelman et al. \(2017\)](#) proposes to use:

$$R^2_{Gelman} = \frac{\text{Var}(\sum_{i=1}^n \hat{y}_i^s)}{\text{Var}(\sum_{i=1}^n \hat{y}_i^s) + \text{Var}(\sum_{i=1}^n e_i^s)}, \quad i = 1, \dots, n, \quad (2.8)$$

, where $\hat{y}_i^s = E(y | X_i, \theta^s)$ 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 formula 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](#)). However, we can no longer interpret an increase in R^2 as a improved fit to a fixed target because the denominator of R^2 is no longer fixed . This 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 fullmodel and the covariance matrix of the predictors. We therefore compare all submodels of a posterior sample to the same fixed value. How it is possible to get the R^2 of the submodels from the full model is shown in the next section.

The variance of the linear model can be written as

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

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

Writting it this way makes it clear that the predictors in (2.8) could also be taken as random ([Gelman et al., 2017](#)). The predictors are then called stochastic predictors. We have then an additional uncertainty in the R^2 formula that can have a large influence on the R^2 values.

In the Bayesian framework the σ^2 parameter is explicitly modeled in the standard linear regression setting. It is therefor possible to sample the σ^2 parameter from its posterior distribution instead of defining the error as in (2.8), which woud lead to the following definition:

$$R^2 = \frac{\text{Var}(\sum_{i=1}^n \hat{y}_i^s)}{\text{Var}(\sum_{i=1}^n \hat{y}_i^s) + \sigma^2} \quad (2.10)$$

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

In practice (2.10) and (2.8) should lead to similar values in the standard linear model. In my opinion it is more reasonable to go the full bayesian route and sample σ^2 . This provides us 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 (2.10). The benefit of (2.8) is that it also works for generalized linear models, where we often have no separate variance parameter.

As a side note: Instead of using $V(\sum_{n=1}^N e_n^s)$ we could also use $\sum(y - \hat{y})^2/(n - 1)$ 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 the samples of the posterior parameters are used instead, the mean of the residuals is not 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 exactly 0 when the samples of the posteriors are used for the regression coefficients. When we use $\text{var}(y_i - \hat{y}_i)$, we do not include this uncertainty and fix the mean to 0. In my understanding, because we do not estimate the mean and take the posterior sample as the possible true regression parameter we would not need to divide by $n - 1$ but only by n . Because this is also true for the fixed effects it does not matter if we divide by $n - 1$ or by n as long as we use the same for the fixed effects and the error term.

For two predictors (2.9) simplifies to

$$\text{Var}(y) = \beta_1^2 \text{Var}(X_1) + 2\beta_1\beta_2 \text{Cov}(X_1, X_2) + \beta_2^2 \text{Var}(X_2) + \sigma^2, \quad (2.11)$$

When predictor X_1 is alone in the model the explained variance includes the variance of the predictor itself, the whole covariance term and in addition some of the contribution of the variance of X_2 in (2.11) . In mathematical notation that is

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

The contribution of the second regressor is then simply the difference to the total explained variance.

In the general case with p regressors, the conditional variance formula (2.12) can be used to calculate the R^2 of all submodels. The conditional variance formula can for example 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

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

and the conditional variance is

$$\text{Var}(\mathbf{Y}_1|\mathbf{y}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \quad (2.12)$$

Lets assume we are interested in the R^2 of a submodel containing the predictors $\mathbf{X}_{q...p}$, and regression coefficients $\boldsymbol{\beta}^\top = (\beta_1, \dots, \beta_p)$ without the intercept. The regression coefficients are further separated in $\boldsymbol{\beta}_{1,...,q-1}^\top = (\beta_1, \dots, \beta_{q-1})$ and $\boldsymbol{\beta}_{q,...,p}^\top = (\beta_q, \dots, \beta_p)$.

As in the normal distribution example above we have the covariance matrix of p predictors written as

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

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

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

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

The total explained variance of the model including predictors $\mathbf{X}_{1...p}$ omits simply the σ^2 parameter in equation , which is

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

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

$$\text{evar}(\mathbf{X}_{q...p}) = \text{evar}(\mathbf{X}_{1,...,p}) - \boldsymbol{\beta}_{1,...,q-1}^\top \text{Cov}(\mathbf{X}_{1,...,q-1} | \mathbf{x}_{q...p}) \boldsymbol{\beta}_{1,...,q-1}. \quad (2.13)$$

To gain the the R^2 value of the submodel we need to divide the explained variance by the total variance,

$$\text{evar}(\mathbf{X}_{q,...,p}) / \text{Var}(\mathbf{Y}). \quad (2.14)$$

The LMG formula requires calculation of the R^2 values for all $2^p - 1$ submodels.

In the Bayesian setting we do have a whole probability distribution for each regression parameter. We can sample the regression parameters from the posterior joint distribution of the fullmodel and use the conditional variance formula to calculate the explained variance of all submodels for each parameter sample. As Gelman notes their R^2 can no longer be interpreted as a fit to a fixed target. For the LMG formula this may be problematic. However, using the conditional variance formula to calculate the R^2 of the submodels, the same total variance is used for a sample of the joint posterior distribution. The important property that all shares should be non-negative and the dependence of the submodels to each other is then respected for

each sample. With dependence, i mean the interconnection of the R^2 values of the submodels. As an example for a violation of this interconnection lets assume we have uncorrelated predictors. Instead of fitting the full model and use 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. The LMG values could then be built by sampling a parameter from each submodel. The problem is then that the parameter values change in each submodel, even if the predictors are uncorrelated. 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. A further of fitting one full model only is that we only need to fit one Bayesian model including all predictors. This makes it possible to calculate the LMG values also in the Bayesian framework in a reasonable amount of time.

2.1 Bayesian Regression

This Section provides a short introduction Bayesian regression and about some assumptions. It is summarized from the book (Bayesian Analysis for the Social Sciences, 2009). In regression analysis we are interested in the dependence of \mathbf{y} on \mathbf{X} . The conditional mean of a continuous response variable $\mathbf{y} = (y_1, \dots, y_n)^\top$ is related to a $n \times k$ predictor matrix \mathbf{X} via a linear model,

$$\mathbb{E}(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}, \quad (2.15)$$

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 can be written as

$$\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \quad (2.16)$$

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

$$p(\mathbf{y}, \mathbf{X} | \boldsymbol{\theta}) = p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_{y|x}) p(\mathbf{X} | \boldsymbol{\theta}_x), \quad (2.17)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x)^\top$. The weak exogeneity assumption implicates that the whole information about \mathbf{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 \mathbf{y}_i . The interest of regression is mostly in the posterior parameters $\boldsymbol{\theta}_{y|x}$. These posterior densities are proportional to likelihood of the data multiplied by the prior density. The joint density $p(\mathbf{y}, \mathbf{X} | \boldsymbol{\theta})$ is used to learn about the posterior parameters, via Bayes Rule

$$p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}, \mathbf{X} | \boldsymbol{\theta}) p(\boldsymbol{\theta}). \quad (2.18)$$

The dependence of \mathbf{y} on \mathbf{X} is captured in the parameters $\boldsymbol{\theta}_{y|x} = (\boldsymbol{\beta}, \sigma^2)$. 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

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}_x | \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2)}{p(\mathbf{y} | \mathbf{X})} \times \frac{p(\mathbf{X} | \boldsymbol{\theta}_x) p(\boldsymbol{\theta}_x)}{p(\mathbf{X})} \quad (2.19)$$

This factorization shows that under the above mentioned assumptions the posterior inference about the parameters $\boldsymbol{\theta}_{y|x} = (\boldsymbol{\beta}, \sigma^2)$ is independent from the inference about $\boldsymbol{\theta}_x$ given data

\mathbf{X} . This also means that the assumptions about \mathbf{X} being fixed or stochastic result in the same posterior density of $\boldsymbol{\theta}_{y|x}$. In the case of fixed 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 $\boldsymbol{\theta}_{y|x} = (\boldsymbol{\beta}, \sigma^2)$ from knowing $\boldsymbol{\theta}_x$.

In regression the focus is on $\boldsymbol{\theta}_{y|x} = (\boldsymbol{\beta}, \sigma^2)$, for which under some assumptions it does not matter whether we assume fixed or stochastic predictors. For the LMG formula the variance of the predictors is also incorporated. The LMG formula may be especially interesting for continuous predictors, which in most cases are stochastic. For stochastic predictors the information about $\boldsymbol{\theta}_x$ would therefore also be relevant. As seen in equation ... inference about $\boldsymbol{\theta}_x$ is independent from inference about $\boldsymbol{\theta}_{y|x}$. If we have stochastic predictors and ignore dependence we just use an estimate of the covariance matrix and do not incorporate this uncertainty. 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 we know the distribution of the $p(\mathbf{X})$ the $\boldsymbol{\theta}_x$ could be estimated. However, the computation times are then much higher. We would need to do the whole LMG calculation for each posterior covariance sample of the predictors. Depending on the number of predictors this would quite take some time. In most cases the problem is that we do not know the distribution of the \mathbf{X} . As a practical solution we could then use nonparametric bootstrapping to include the uncertainty of the stochastic predictors in the LMG formula. We would then also need to do the whole LMG calculations for each bootstrap sample of the covariance matrix. There exist also different covariance estimator. The shrinkage method may be an interesting estimator with some interesting properties.

Chapter 3

Examples

In the following section the Bayesian LMG implementation is presented on two examples. The first examples simulates data, the second examples uses real data.

3.1 Simulated Data

Lets assume a simple model:

$$Y_i \sim \mathcal{N}(\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4, \sigma^2), \quad (3.1)$$

$$\beta_1 = 0.5, \beta_2 = 1, \beta_3 = 2, \beta_4 = 0, \sigma^2 = 1 \quad (3.2)$$

$$\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4 \sim \mathcal{N}(0, 1) \quad (3.3)$$

The values of the four predictors are sampled from a standard normal distribution. These values are then multiplied by the regression coefficients to obtain the dependent variable. A standard normal distributed error is added. Fifty observations were sampled.

The following Code was used to simulate the data :

```
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
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)
```

The model is fitted using the rstanarm package with the default priors for the regression and σ^2 parameter. For computational reasons a small burning period of 1000 and a sample size of 1000 were chosen. For each posterior sample of the parameters the R^2 value is calculated. The R^2 of the submodels is then calculated by the conditional variance formula for each posterior sample.

```
post2 <- stan_glm(y ~ 1 + x1 + x2 + x3 + x4,
                     data = df,
                     chains = 2, cores = 1)

## 
## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
```

```

## 
## Gradient evaluation took 0.000122 seconds
## 1000 transitions using 10 leapfrog steps per transition would take 1.22 seconds.
## Adjust your expectations accordingly!
##
##
## Iteration: 1 / 2000 [ 0%] (Warmup)
## Iteration: 200 / 2000 [ 10%] (Warmup)
## Iteration: 400 / 2000 [ 20%] (Warmup)
## Iteration: 600 / 2000 [ 30%] (Warmup)
## Iteration: 800 / 2000 [ 40%] (Warmup)
## Iteration: 1000 / 2000 [ 50%] (Warmup)
## Iteration: 1001 / 2000 [ 50%] (Sampling)
## Iteration: 1200 / 2000 [ 60%] (Sampling)
## Iteration: 1400 / 2000 [ 70%] (Sampling)
## Iteration: 1600 / 2000 [ 80%] (Sampling)
## Iteration: 1800 / 2000 [ 90%] (Sampling)
## Iteration: 2000 / 2000 [100%] (Sampling)
##
## Elapsed Time: 0.074651 seconds (Warm-up)
##                 0.079635 seconds (Sampling)
##                 0.154286 seconds (Total)
##
##
## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
##
## Gradient evaluation took 1.6e-05 seconds
## 1000 transitions using 10 leapfrog steps per transition would take 0.16 seconds.
## Adjust your expectations accordingly!
##
##
## Iteration: 1 / 2000 [ 0%] (Warmup)
## Iteration: 200 / 2000 [ 10%] (Warmup)
## Iteration: 400 / 2000 [ 20%] (Warmup)
## Iteration: 600 / 2000 [ 30%] (Warmup)
## Iteration: 800 / 2000 [ 40%] (Warmup)
## Iteration: 1000 / 2000 [ 50%] (Warmup)
## Iteration: 1001 / 2000 [ 50%] (Sampling)
## Iteration: 1200 / 2000 [ 60%] (Sampling)
## Iteration: 1400 / 2000 [ 70%] (Sampling)
## Iteration: 1600 / 2000 [ 80%] (Sampling)
## Iteration: 1800 / 2000 [ 90%] (Sampling)
## Iteration: 2000 / 2000 [100%] (Sampling)
##
## Elapsed Time: 0.07518 seconds (Warm-up)
##                 0.065774 seconds (Sampling)
##                 0.140954 seconds (Total)

#posterior sample
post.sample <- as.matrix(post2)

```

```
#example of the first 10 posterior samples
post.sample[1:10,]

##          parameters
## iterations (Intercept)    x1     x2     x3     x4 sigma
## [1,]      -0.0374 0.857 0.875 0.907 1.12 0.875
## [2,]       0.3085 1.185 0.569 1.243 1.20 0.958
## [3,]      -0.0216 0.916 0.522 1.130 1.19 1.159
## [4,]      -0.1909 1.104 0.721 0.860 1.20 0.749
## [5,]       0.0495 1.099 0.750 0.815 1.32 0.795
## [6,]      -0.0045 1.215 0.791 0.899 1.23 0.801
## [7,]      -0.0487 1.012 0.763 0.860 1.10 0.839
## [8,]       0.0875 1.186 0.951 0.901 1.22 0.834
## [9,]       0.1466 1.013 0.804 1.019 1.24 0.791
## [10,]      0.1429 1.015 0.930 1.098 1.31 0.900

#no need for the intercept, last parameter is sigma
post.sample <- post.sample[,-1]

#data frame with all submodels
df.rtwos <- rtwos(df[,2:5], post.sample)

df.rtwos[,1:5]

##          X1     X2     X3     X4     X5
## none      0.0000 0.0000 0.0000 0.0000 0.0000
## x1        0.0888 0.1559 0.0948 0.1604 0.1412
## x2        0.1310 0.0398 0.0368 0.0777 0.0755
## x3        0.3557 0.4412 0.3882 0.3359 0.2968
## x4        0.2861 0.2612 0.2809 0.3244 0.3690
## x1 x2     0.2438 0.2136 0.1449 0.2631 0.2398
## x1 x3     0.4148 0.5527 0.4509 0.4567 0.4031
## x1 x4     0.4110 0.4622 0.4125 0.5359 0.5616
## x2 x3     0.4416 0.4559 0.4023 0.3806 0.3415
## x2 x4     0.4504 0.3190 0.3358 0.4299 0.4739
## x3 x4     0.5854 0.6426 0.6108 0.6019 0.6073
## x1 x2 x3  0.5187 0.5782 0.4727 0.5202 0.4654
## x1 x2 x4  0.6103 0.5468 0.4884 0.6780 0.7013
## x1 x3 x4  0.6741 0.7914 0.7033 0.7675 0.7593
## x2 x3 x4  0.6999 0.6692 0.6371 0.6694 0.6769
## all       0.8170 0.8365 0.7440 0.8651 0.8581
```

After the R^2 for each posterior sample and their corresponding submodels is calculated, the package hier.part is used to calculate the LMG value for each posterior sample.

```
# prepare data frame for LMG values

LMG.Vals.I<-matrix(0, 4, dim(df.rtwos)[2])

LMG.Vals.J<-matrix(0, 4, dim(df.rtwos)[2])
```

```

LMG.Vals.T<-matrix(0, 4, dim(df.rtwos)[2])

for(i in 1:dim(df.rtwos)[2]){

gofn<-df.rtwos[,i]

obj.Gelman<-partition(gofn, pcan = 4, var.names = names(df[,2:5]))

LMG.Vals.I[,i]=obj.Gelman$IJ[,1]
LMG.Vals.J[,i]=obj.Gelman$IJ[,2]
LMG.Vals.T[,i]=obj.Gelman$IJ[,3]
}

varnames <- row.names(obj.Gelman$IJ)

# posterior LMG distribution of each variable
quantile(LMG.Vals.I[1,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.0867 0.1641 0.2472

quantile(LMG.Vals.I[2,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.0571 0.1186 0.2024

quantile(LMG.Vals.I[3,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.165 0.258 0.354

quantile(LMG.Vals.I[4,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.186 0.285 0.383

quantile(LMG.Vals.J[1,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## -0.0253 -0.0173 -0.0091

quantile(LMG.Vals.J[2,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## -0.01655 -0.00941 -0.00269

quantile(LMG.Vals.J[3,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.0585 0.0703 0.0777

```

```

quantile(LMG.Vals.J[4,], c(0.025, 0.5, 0.975))

##      2.5%      50%    97.5%
## -0.02031 -0.01092 -0.00092

quantile(LMG.Vals.T[1,], c(0.025, 0.5, 0.975))

##      2.5%      50%    97.5%
## 0.0736 0.1462 0.2270

quantile(LMG.Vals.T[2,], c(0.025, 0.5, 0.975))

##      2.5%      50%    97.5%
## 0.0497 0.1092 0.1911

quantile(LMG.Vals.T[3,], c(0.025, 0.5, 0.975))

##      2.5%      50%    97.5%
## 0.224 0.329 0.429

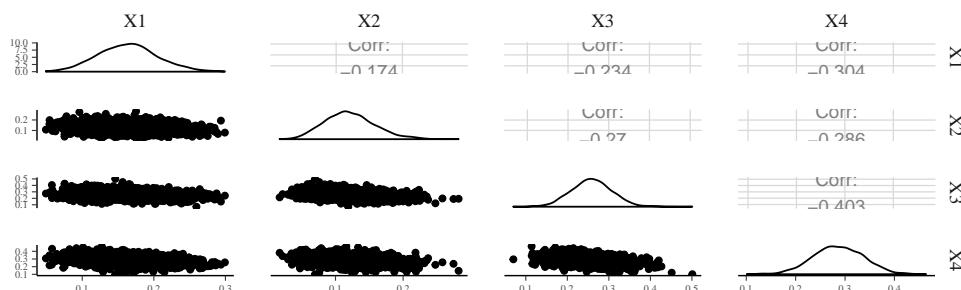
quantile(LMG.Vals.T[4,], c(0.025, 0.5, 0.975))

##      2.5%      50%    97.5%
## 0.177 0.275 0.373

# some example how it could be displayed
dat <- data.frame(t(LMG.Vals.I))

pairs.chart <- ggpairs(dat, lower = list(list(combo = "facetdensity")),
upper = list(list(combo = "jitter")))
pairs.chart

```



```

# Comparison to relaimpo package

fit <- lm(y~., data=df)

##### compare to relimp package

run<-boot.relimp(fit, fixed=TRUE)

booteval.relimp(run, bty = "perc", level = 0.95,
                 sort = FALSE, norank = TRUE, nodiff = TRUE,
                 typesel = c("lmg"))

```

```

## Response variable: y
## Total response variance: 4.36
## Analysis based on 50 observations
##
## 4 Regressors:
## x1 x2 x3 x4
## Proportion of variance explained by model: 84.8%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##      lmg
## x1 0.167
## x2 0.121
## x3 0.266
## x4 0.294
##
## Average coefficients for different model sizes:
##
##      1X   2Xs   3Xs   4Xs
## x1 0.968 1.004 1.046 1.097
## x2 0.773 0.789 0.815 0.855
## x3 1.160 1.087 0.997 0.882
## x4 1.091 1.106 1.126 1.153
##
## Confidence interval information ( 1000 bootstrap replicates, bty= perc ):
## Relative Contributions with confidence intervals:
##
##           Lower   Upper
## percentage 0.95   0.95
## x1.lmg 0.1669 0.0983 0.2576
## x2.lmg 0.1211 0.0631 0.2042
## x3.lmg 0.2659 0.1810 0.3666
## x4.lmg 0.2941 0.2075 0.3899
##
## CAUTION: Bootstrap confidence intervals can be somewhat liberal.
## NOTE: X-matrix has been considered as fixed for bootstrapping.

```

Variable	I	J	Total
x1	0.164 (0.087, 0.247)	-0.017 (-0.025, -0.009)	0.146 (0.074, 0.227)
x2	0.119 (0.057, 0.202)	-0.009 (-0.017, -0.003)	0.109 (0.05, 0.191)
x3	0.258 (0.165, 0.354)	0.07 (0.059, 0.078)	0.329 (0.224, 0.429)
x4	0.285 (0.186, 0.383)	-0.011 (-0.02, -9.196 × 10 ⁻⁴)	0.275 (0.177, 0.373)

Table 3.1: Variance decomposition for stochastic predictors using bootstrap. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Using the default uninformative priors, the LMG distributions obtained from the Bayesian framework are very similar to the bootstrap confidence intervals of the LMG estimates obtained

from the relaimpo package. In both cases fixed regressors are assumed. In the example above the predictors were sampled from a normal distribution. It would therefore be more reasonable to assume stochastic predictors. As noted in ... under the assumption of weak exogeneity and conditional independence the posterior distributions of the regression parameters β are valid for fixed and stochastic predictors. Inference about the covariance matrix can be seen as an independent problem. G recommends in most cases to use the non fixed regressor option when calculating bootstrap confidence intervals. The confidence intervals will then in general be a bit larger. If we want to include this uncertainty in the Bayesian framework, we would need some ideas about the distribution of the predictor variables \mathbf{X} . It is then possible to obtain posterior distributions of their corresponding covariance matrix. As a practical solution nonparametric bootstrap may be used to include the uncertainty of the covariance matrix. The following code includes the uncertainty of the stochastic predictors.

```
#Code assuming stochastic predictors
run.stochastic<-boot.relimp(fit, fixed=FALSE)

booteval.relimp(run.stochastic, bty = "perc", level = 0.95,
                 sort = FALSE, norank = TRUE, nodiff = TRUE,
                 typesel = c("lmg"))

## Response variable: y
## Total response variance: 4.36
## Analysis based on 50 observations
##
## 4 Regressors:
## x1 x2 x3 x4
## Proportion of variance explained by model: 84.8%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##      lmg
## x1 0.167
## x2 0.121
## x3 0.266
## x4 0.294
##
## Average coefficients for different model sizes:
##
##      1X   2Xs   3Xs   4Xs
## x1 0.968 1.004 1.046 1.097
## x2 0.773 0.789 0.815 0.855
## x3 1.160 1.087 0.997 0.882
## x4 1.091 1.106 1.126 1.153
##
## Confidence interval information ( 1000 bootstrap replicates, bty= perc ):
## Relative Contributions with confidence intervals:
##
##           Lower   Upper
## percentage 0.95   0.95
## x1.lmg    0.1669  0.0883 0.2838
```

```

## x2.lmg 0.1211      0.0468 0.2353
## x3.lmg 0.2659      0.1346 0.4001
## x4.lmg 0.2941      0.1493 0.4099
##
## CAUTION: Bootstrap confidence intervals can be somewhat liberal.

#-----


df.rtwos.boot <-rtwos.boot(df[,2:5], post.sample, 100)

n.boot = 100

LMG.Vals.I.boot<-array(0, c(4,dim(df.rtwos.boot)[2], n.boot))
LMG.Vals.J.boot<-array(0, c(4,dim(df.rtwos.boot)[2], n.boot))
LMG.Vals.T.boot<-array(0, c(4,dim(df.rtwos.boot)[2], n.boot))

for (b in 1:n.boot){

  for(i in 1:dim(df.rtwos.boot)[2]){

    gofn<-df.rtwos.boot[,i,b]

    obj.Gelman<-partition(gofn, pcan = 4, var.names = names(df[,2:5]))

    LMG.Vals.I.boot[,i, b]=obj.Gelman$IJ[,1]
    LMG.Vals.J.boot[,i, b]=obj.Gelman$IJ[,2]
    LMG.Vals.T.boot[,i, b]=obj.Gelman$IJ[,3]

  }

}

quantile(c(LMG.Vals.I.boot[1,1:1000,]), c(0.025, 0.5, 0.975))

##   2.5%   50% 97.5%
## 0.0724 0.1652 0.2766

quantile(c(LMG.Vals.I.boot[2,1:1000,]), c(0.025, 0.5, 0.975))

##   2.5%   50% 97.5%
## 0.0369 0.1142 0.2263

quantile(c(LMG.Vals.I.boot[3,1:1000,]), c(0.025, 0.5, 0.975))

##   2.5%   50% 97.5%
## 0.140 0.261 0.380

quantile(c(LMG.Vals.I.boot[4,1:1000,]), c(0.025, 0.5, 0.975))

##   2.5%   50% 97.5%
## 0.154 0.278 0.408

```

```
#very similar values as in the confidence intervals for stochastic predictors

#what if we have prior knowledge
my_prior <- normal(location = c(1, 1, 1, 1), scale = c(0.01, 0.01, 0.01, 0.01), autoscale = FALSE)

post2 <- stan_glm(y ~ 1 + x1 + x2 + x3 + x4,
                    data = df, prior = my_prior,
                    chains = 2, cores = 1)

##  

## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).  

##  

## Gradient evaluation took 2.6e-05 seconds  

## 1000 transitions using 10 leapfrog steps per transition would take 0.26 seconds.  

## Adjust your expectations accordingly!  

##  

##  

## Iteration: 1 / 2000 [ 0%] (Warmup)  

## Iteration: 200 / 2000 [ 10%] (Warmup)  

## Iteration: 400 / 2000 [ 20%] (Warmup)  

## Iteration: 600 / 2000 [ 30%] (Warmup)  

## Iteration: 800 / 2000 [ 40%] (Warmup)  

## Iteration: 1000 / 2000 [ 50%] (Warmup)  

## Iteration: 1001 / 2000 [ 50%] (Sampling)  

## Iteration: 1200 / 2000 [ 60%] (Sampling)  

## Iteration: 1400 / 2000 [ 70%] (Sampling)  

## Iteration: 1600 / 2000 [ 80%] (Sampling)  

## Iteration: 1800 / 2000 [ 90%] (Sampling)  

## Iteration: 2000 / 2000 [100%] (Sampling)  

##  

## Elapsed Time: 0.087333 seconds (Warm-up)  

##                 0.064388 seconds (Sampling)  

##                 0.151721 seconds (Total)  

##  

##  

## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).  

##  

## Gradient evaluation took 1.5e-05 seconds  

## 1000 transitions using 10 leapfrog steps per transition would take 0.15 seconds.  

## Adjust your expectations accordingly!  

##  

##  

## Iteration: 1 / 2000 [ 0%] (Warmup)  

## Iteration: 200 / 2000 [ 10%] (Warmup)  

## Iteration: 400 / 2000 [ 20%] (Warmup)  

## Iteration: 600 / 2000 [ 30%] (Warmup)  

## Iteration: 800 / 2000 [ 40%] (Warmup)  

## Iteration: 1000 / 2000 [ 50%] (Warmup)  

## Iteration: 1001 / 2000 [ 50%] (Sampling)  

## Iteration: 1200 / 2000 [ 60%] (Sampling)
```

```

## Iteration: 1400 / 2000 [ 70%] (Sampling)
## Iteration: 1600 / 2000 [ 80%] (Sampling)
## Iteration: 1800 / 2000 [ 90%] (Sampling)
## Iteration: 2000 / 2000 [100%] (Sampling)
##
## Elapsed Time: 0.089243 seconds (Warm-up)
##                  0.063377 seconds (Sampling)
##                  0.15262 seconds (Total)

#posterior sample
post.sample <- as.matrix(post2)

#example of the first 10 posterior samples
post.sample[1:10,]

##          parameters
## iterations (Intercept)      x1      x2      x3      x4 sigma
## [1,]      0.0626  0.988  0.999  1.005  0.997  0.766
## [2,]      0.1738  1.009  1.006  0.995  1.002  0.853
## [3,]     -0.0564  0.993  0.990  1.012  0.993  0.926
## [4,]     -0.1099  0.997  0.999  1.018  0.998  1.143
## [5,]      0.1825  1.003  0.998  0.979  1.004  0.759
## [6,]     -0.1708  0.999  1.017  1.009  1.000  0.906
## [7,]      0.1113  1.006  1.000  0.990  0.991  0.770
## [8,]     -0.0681  0.991  1.001  1.008  1.011  0.952
## [9,]      0.0660  1.010  0.995  0.989  1.011  0.764
## [10,]     -0.0335  0.991  1.004  1.010  0.991  0.938

#no need for the intercept, last parameter is sigma
post.sample <- post.sample[,-1]

#data frame with all submodels

df.rtwos.boot <- rtwos.boot(df[,2:5], post.sample, 10)

n.boot = 10

LMG.Vals.I.boot.p<-array(0, c(4,dim(df.rtwos.boot)[2], n.boot))

for (b in 1:n.boot){

  for(i in 1:dim(df.rtwos.boot)[2]){

    gofn<-df.rtwos.boot[,i,b]

    obj.Gelman<-partition(gofn, pcan = 4, var.names = names(df[,2:5]))

    LMG.Vals.I.boot.p[,i, b]=obj.Gelman$IJ[,1]
  }
}

```

```

}

quantile(c(LMG.Vals.I.boot.p[1,1:1000,]), c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.0798 0.1373 0.1899

quantile(c(LMG.Vals.I.boot.p[2,1:1000,]), c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.105 0.177 0.217

quantile(c(LMG.Vals.I.boot.p[3,1:1000,]), c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.277 0.324 0.380

quantile(c(LMG.Vals.I.boot.p[4,1:1000,]), c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.161 0.213 0.269

```

Variable	I	J	Total
x1	0.165 (0.072, 0.277)	-0.01 (-0.08, 0.1)	0.154 (0.03, 0.326)
x2	0.114 (0.037, 0.226)	-0.014 (-0.063, 0.078)	0.102 (0.005, 0.268)
x3	0.261 (0.14, 0.38)	0.073 (-0.026, 0.207)	0.336 (0.145, 0.536)
x4	0.278 (0.154, 0.408)	-0.007 (-0.093, 0.133)	0.274 (0.095, 0.474)

Table 3.2: Variance decomposition. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

In the following example we know that the **X** are coming from a normal distribution. The covariance matrix of the distribution is estimated in a Bayesian way. The package Jags is used.

```

zy = df[,2:5]

#-----
# The rest can remain unchanged, except for the specification of difference of
# correlations at the very end.
#-----

# Load some functions used below:
# Install the ellipse package if not already:
# Standardize the data:

#zy = apply(y, 2, function(yVec){(yVec-mean(yVec))/sd(yVec)})
# Assemble data for sending to JAGS:
dataList = list(
  zy = zy ,
  Ntotal = nrow(zy) ,
  Nvar = ncol(zy) ,
  # Include original data info for transforming to original scale:

```

```

# For wishart (dwish) prior on inverse covariance matrix:
zRscal = ncol(zy) , # for dwish prior
zRmat = diag(x=1,nrow=ncol(zy)) # Rmat = diag(apply(y,2,var))
)

# Define the model:
modelString =
model {
for ( i in 1:Ntotal ) {
zy[i,1:Nvar] ~ dmnorm( zMu[1:Nvar] , zInvCovMat[1:Nvar,1:Nvar] )
}
for ( varIdx in 1:Nvar ) { zMu[varIdx] ~ dnorm( 0 , 1/2^2 ) }
zInvCovMat ~ dwish( zRmat[1:Nvar,1:Nvar] , zRscal )
# Convert invCovMat to sd and correlation:
zCovMat <- inverse( zInvCovMat )

}
" # close quote for modelString
writeLines( modelString , con="Jags-MultivariateNormal-model.txt" )

# Run the chains:
nChain = 3
nAdapt = 500
nBurnIn = 500
nThin = 10
nStepToSave = 20000
require(rjags)
jagsModel = jags.model( file="Jags-MultivariateNormal-model.txt" ,

## Compiling model graph
## Resolving undeclared variables
## Allocating nodes
## Graph information:
## Observed stochastic nodes: 50
## Unobserved stochastic nodes: 5
## Total graph size: 82
##
## Initializing model

update( jagsModel , n.iter=nBurnIn )
codaSamples = coda.samples( jagsModel ,
# Convergence diagnostics:
parameterNames = varnames(codaSamples) # get all parameter names

# Examine the posterior distribution:

```

```

mcmcMat = as.matrix(codaSamples)
chainLength = nrow(mcmcMat)

covMat <- array(NA, c(4,4,chainLength))

for (i in 1:chainLength){
covMat[1:4,1:4,i]<-matrix(mcmcMat[i,], 4, 4)
}

covMat <- covMat[1:4,1:4,sample(1:20000, replace=F)]

df.rtwos <-rtwos.covm(df[,2:5], post.sample, covMat, 10)

n.boot = 10

LMG.Vals.I.covm<-array(0, c(4,dim(df.rtwos)[2], n.boot))
LMG.Vals.J.covm<-array(0, c(4,dim(df.rtwos.boot)[2], n.boot))
LMG.Vals.T.covm<-array(0, c(4,dim(df.rtwos.boot)[2], n.boot))

for (b in 1:n.boot){

  for(i in 1:dim(df.rtwos)[2]){

    gofn<-df.rtwos[,i,b]

    obj.Gelman<-partition(gofn, pcan = 4, var.names = names(df[,2:5]))

    LMG.Vals.I.covm[,i, b]=obj.Gelman$IJ[,1]
    LMG.Vals.J.covm[,i, b]=obj.Gelman$IJ[,2]
    LMG.Vals.T.covm[,i, b]=obj.Gelman$IJ[,3]

  }

}

quantile(c(LMG.Vals.I.covm[1,1:1000,]), c(0.025, 0.5, 0.975))

##   2.5%     50% 97.5%
## 0.0893 0.1507 0.2076

quantile(c(LMG.Vals.I.covm[2,1:1000,]), c(0.025, 0.5, 0.975))

##   2.5%     50% 97.5%
## 0.104 0.141 0.222

quantile(c(LMG.Vals.I.covm[3,1:1000,]), c(0.025, 0.5, 0.975))

##   2.5%     50% 97.5%
## 0.264 0.305 0.452

```

```
quantile(c(LMG.Vals.I.covm[4,1:1000,]), c(0.025, 0.5, 0.975))

## 2.5%   50% 97.5%
## 0.141 0.204 0.315
```

Variable	I	J	Total
x1	0.151 (0.089, 0.208)	0.016 (-0.063, 0.114)	0.173 (0.03, 0.322)
x2	0.141 (0.104, 0.222)	-0.008 (-0.091, 0.04)	0.137 (0.038, 0.246)
x3	0.305 (0.264, 0.452)	0.118 (0.04, 0.167)	0.432 (0.32, 0.564)
x4	0.204 (0.141, 0.315)	0.019 (-0.047, 0.133)	0.22 (0.096, 0.448)

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

Using the bootstrap samples of the covariance matrix or samples from the posterior covariance matrix resulted in very similar LMG values. 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 of the covariance matrix. For stochastic predictors the uncertainty about the covariance matrix is reflected in the large credible intervals. Even when we would knew the exact regression parameters, there is alot of uncertainty in the LMG values caused by the uncertainty about the covariance matrix. Code xx shows the uncertainty about the LMG values caused by the uncertainty about the covariance matrix.

Another interesting option in some cases would be to use the shrinkage estimate of the covariance matrix.

#How much variance is effectively in the bootstrap matrix when we know the regression param

```
#fake post sample

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
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)

post.sample <- matrix(1, 1, 5)

df.rtwos.boot <- rtwos.boot(df[,2:5], post.sample, 1000)

n.boot = 1000

LMG.Vals.I.boot<-array(0, c(4, dim(df.rtwos.boot)[2], n.boot))

for (b in 1:n.boot){

  for(i in 1:dim(df.rtwos.boot)[2]){

    df.rtwos.boot[i,b] =
```

```

gofn<-df.rtwos.boot[,i,b]

obj.Gelman<-partition(gofn, pcan = 4, var.names = names(df[,2:5]))

LMG.Vals.I.boot[,i, b]=obj.Gelman$IJ[,1]
}

}

quantile(c(LMG.Vals.I.boot[1,1,]), c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.0809 0.1493 0.2391

quantile(c(LMG.Vals.I.boot[2,1,]), c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.162 0.252 0.334

quantile(c(LMG.Vals.I.boot[3,1,]), c(0.025, 0.5, 0.975))

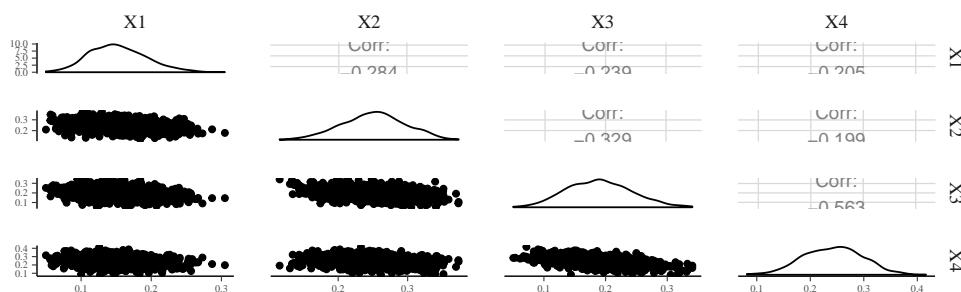
## 2.5% 50% 97.5%
## 0.0944 0.1888 0.2925

quantile(c(LMG.Vals.I.boot[4,1,]), c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.137 0.245 0.347

dat <- data.frame(t(LMG.Vals.I.boot[1:4,1,]))
pairs.chart <- ggpairs(dat, lower = list(list(combo = "facetdensity")),
upper = list(list(combo = "jitter")))
pairs.chart

```



```

cov(df[,2:5])

##           x1      x2      x3      x4
## x1  0.8905 -0.0969  0.0377  0.1193
## x2 -0.0969  1.4187  0.0778  0.2580
## x3  0.0377  0.0778  1.0442  0.0341
## x4  0.1193  0.2580  0.0341  1.1203

```

```
#Comparison of sample covariance and shrink covariance matrix

cov(df[,2:5])

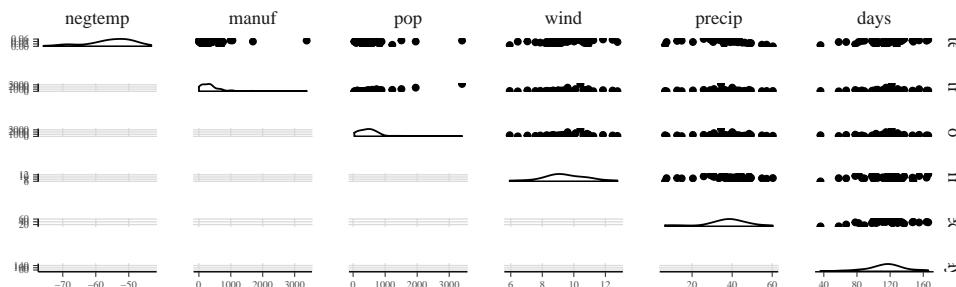
##          x1      x2      x3      x4
## x1  0.8905 -0.0969  0.0377  0.1193
## x2 -0.0969  1.4187  0.0778  0.2580
## x3  0.0377  0.0778  1.0442  0.0341
## x4  0.1193  0.2580  0.0341  1.1203

cov.shrink(df[,2:5])

## Estimating optimal shrinkage intensity lambda.var (variance vector): 1
##
## Estimating optimal shrinkage intensity lambda (correlation matrix): 1
##          x1      x2      x3      x4
## x1  1.08  0.00  0.00  0.00
## x2  0.00  1.08  0.00  0.00
## x3  0.00  0.00  1.08  0.00
## x4  0.00  0.00  0.00  1.08
## attr(,"lambda")
## [1] 1
## attr(,"lambda.estimated")
## [1] TRUE
## attr(,"class")
## [1] "shrinkage"
## attr(,"lambda.var")
## [1] 1
## attr(,"lambda.var.estimated")
## [1] TRUE
```

3.2 Real World Data

The following example data are taken from the book Bayesian Regression Modeling with INLA. The data were about air pollution in 41 cities in the United States originally published in Everitt (2006). The data consists of the SO2 level as the dependent variable and six explanatory variables. Two of the explanatory variables are related to human ecology (pop, manuf) and four others are related to climate (negtemp, wind, precip, days).



A simple linear regression model including the dependent variable SO2 and the six explanatory variables is fitted with the lm command in R. The R^2 of the full model is 0.670.

Table 3.4: Variable description of air pollution data

Variable Name	Description	Codes/Values
SO2	sulfur dioxide content of air	microgrammes per cubic meter
negtemp	negative value of average	fahrenheit
manuf	number of manufacturing enterprises employing 20 or more workers	integers
pop	population size in thousands (1970 census)	numbers
wind	average wind speed	miles per hour
precip	average annual precipitation	inches
days	average number of days with precipitation per year	integers

```

usair.lm <- lm(SO2~., data = usair)

#Code assuming stochastic predictors
run<-boot.relimp(usair.lm, fixed=TRUE)

booteval.relimp(run, bty = "perc", level = 0.95,
                 sort = FALSE, norank = TRUE, nodiff = TRUE,
                 typesel = c("lmg"))

## Response variable: SO2
## Total response variance: 551
## Analysis based on 41 observations
##
## 6 Regressors:
## negtemp manuf pop wind precip days
## Proportion of variance explained by model: 67%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##          lmg
## negtemp 0.1111
## manuf   0.2965
## pop     0.1661
## wind    0.0167
## precip  0.0187
## days    0.0604
##
## Average coefficients for different model sizes:
##
##          1X      2Xs      3Xs      4Xs      5Xs      6Xs
## negtemp 1.4081  1.3344  1.28118  1.2577  1.2826  1.2679
## manuf   0.0269  0.0373  0.04643  0.0539  0.0598  0.0649
## pop     0.0200  0.0044 -0.00925 -0.0207 -0.0304 -0.0393
## wind    1.5557 -0.0248 -1.38483 -2.3795 -2.9803 -3.1814
## precip  0.1083  0.1147  0.16427  0.2601  0.4057  0.5124
## days    0.3273  0.2979  0.24111  0.1557  0.0425 -0.0521

```

```

##  

##  

##  Confidence interval information ( 1000 bootstrap replicates, bty= perc ):  

##  Relative Contributions with confidence intervals:  

##  

##          Lower   Upper  

##      percentage 0.95  0.95  

## negtemp.lmg 0.1111  0.0399 0.2206  

## manuf.lmg   0.2965  0.2079 0.4126  

## pop.lmg     0.1661  0.1134 0.2443  

## wind.lmg    0.0167  0.0086 0.0718  

## precip.lmg  0.0187  0.0075 0.0676  

## days.lmg    0.0604  0.0187 0.1359  

##  

## CAUTION: Bootstrap confidence intervals can be somewhat liberal.  

## NOTE: X-matrix has been considered as fixed for bootstrapping.  

bayes.usair <- stan_glm(SO2 ~ . ,  

                         data = usair,  

                         chains = 4, cores = 4)  

#posterior sample  

post.sample <- as.matrix(bayes.usair)  

#example of the first 10 posterior samples  

post.sample[1:10,]  

##          parameters  

## iterations (Intercept) negtemp manuf      pop  wind precip      days sigma  

## [1,]      154.7    1.47 0.0610 -0.0321 -5.29  0.310 -0.1161 12.9  

## [2,]      138.4    1.47 0.0510 -0.0286 -4.69  0.199  0.0271 13.0  

## [3,]      159.8    1.71 0.0386 -0.0104 -6.89  0.601 -0.0297 12.8  

## [4,]      135.3    2.07 0.0392 -0.0178 -1.05  1.245 -0.2573 20.0  

## [5,]      220.5    2.51 0.0621 -0.0385 -7.49  1.210 -0.2822 12.2  

## [6,]      151.5    1.64 0.0625 -0.0395 -4.38  0.701 -0.1878 13.2  

## [7,]      95.2     1.13 0.0646 -0.0427 -2.64  0.393  0.0238 13.2  

## [8,]     178.4     2.09 0.0588 -0.0401 -4.25  0.752 -0.1822 16.1  

## [9,]     182.5     1.94 0.0528 -0.0233 -7.38  0.661 -0.0920 15.7  

## [10,]     129.2    1.48 0.0596 -0.0364 -4.12  0.466 -0.0031 11.6  

#no need for the intercept, last parameter is sigma  

post.sample <- post.sample[,-1]  

#data frame with all submodels  

df.rtwos <- rtwos(usair[,2:7], post.sample)  

df.rtwos[,1:3]  

##          X1      X2      X3  

## none      0.00e+00 0.000000 0.000000  

## x1       1.97e-01 0.284523 0.16279  

## x2       4.66e-01 0.347261 0.43399

```

```

## x3          3.00e-01 0.198395 0.30876
## x4          4.83e-05 0.000313 0.00610
## x5          1.10e-02 0.006372 0.00394
## x6          5.92e-02 0.138557 0.14439
## x1 x2      5.68e-01 0.531508 0.51433
## x1 x3      4.68e-01 0.454922 0.44519
## x1 x4      2.27e-01 0.317017 0.21754
## x1 x5      2.02e-01 0.303245 0.21897
## x1 x6      2.01e-01 0.309526 0.21506
## x2 x3      5.91e-01 0.505155 0.49602
## x2 x4      4.96e-01 0.363178 0.49247
## x2 x5      4.73e-01 0.350951 0.44108
## x2 x6      4.90e-01 0.435553 0.52144
## x3 x4      3.15e-01 0.204610 0.34911
## x3 x5      3.08e-01 0.203048 0.31474
## x3 x6      3.48e-01 0.323570 0.43615
## x4 x5      1.11e-02 0.006649 0.00992
## x4 x6      6.14e-02 0.140494 0.16467
## x5 x6      1.27e-01 0.231347 0.16535
## x1 x2 x3  6.28e-01 0.590520 0.53559
## x1 x2 x4  6.54e-01 0.608672 0.63732
## x1 x2 x5  5.70e-01 0.544606 0.55860
## x1 x2 x6  5.69e-01 0.548474 0.55253
## x1 x3 x4  5.50e-01 0.527898 0.56752
## x1 x3 x5  4.74e-01 0.473879 0.50190
## x1 x3 x6  4.71e-01 0.477789 0.49348
## x1 x4 x5  2.37e-01 0.343893 0.29198
## x1 x4 x6  2.31e-01 0.342953 0.27157
## x1 x5 x6  2.02e-01 0.309859 0.22335
## x2 x3 x4  6.28e-01 0.526656 0.56098
## x2 x3 x5  5.97e-01 0.508097 0.50382
## x2 x3 x6  5.94e-01 0.541822 0.55090
## x2 x4 x5  5.03e-01 0.366953 0.49934
## x2 x4 x6  5.29e-01 0.464027 0.60285
## x2 x5 x6  5.25e-01 0.495216 0.52714
## x3 x4 x5  3.24e-01 0.209346 0.35485
## x3 x4 x6  3.75e-01 0.342287 0.50422
## x3 x5 x6  4.02e-01 0.402940 0.44953
## x4 x5 x6  1.33e-01 0.237439 0.19074
## x1 x2 x3 x4 7.01e-01 0.656304 0.65030
## x1 x2 x3 x5 6.28e-01 0.596064 0.57111
## x1 x2 x3 x6 6.28e-01 0.599999 0.56687
## x1 x2 x4 x5 6.61e-01 0.632223 0.70506
## x1 x2 x4 x6 6.54e-01 0.626054 0.67631
## x1 x2 x5 x6 5.71e-01 0.548797 0.56076
## x1 x3 x4 x5 5.63e-01 0.559821 0.65269
## x1 x3 x4 x6 5.52e-01 0.551841 0.61783
## x1 x3 x5 x6 4.74e-01 0.478612 0.50440
## x1 x4 x5 x6 2.38e-01 0.346356 0.29252
## x2 x3 x4 x5 6.34e-01 0.529672 0.56857
## x2 x3 x4 x6 6.35e-01 0.571661 0.63352

```

```

## x2 x3 x5 x6      6.10e-01 0.575817 0.55275
## x2 x4 x5 x6      5.71e-01 0.531670 0.61303
## x3 x4 x5 x6      4.37e-01 0.430562 0.52463
## x1 x2 x3 x4 x5  7.03e-01 0.669594 0.70958
## x1 x2 x3 x4 x6  7.01e-01 0.666749 0.68389
## x1 x2 x3 x5 x6  6.28e-01 0.600015 0.57319
## x1 x2 x4 x5 x6  6.66e-01 0.632442 0.70541
## x1 x3 x4 x5 x6  5.68e-01 0.560186 0.65296
## x2 x3 x4 x5 x6  6.56e-01 0.611735 0.63818
## all                 7.08e-01 0.669885 0.70990

# prepare data frame for LMG values

LMG.Vals.I<-matrix(0, 6, dim(df.rtwos)[2])

for(i in 1:dim(df.rtwos)[2]){

  gofn<-df.rtwos[,i]

  obj.Gelman<-partition(gofn, pcan = 6, var.names = names(usair[,2:7]))

  LMG.Vals.I[,i]=obj.Gelman$IJ[,1]
}

# posterior LMG distribution of each variable
quantile(LMG.Vals.I[1,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.0341 0.1038 0.2061

quantile(LMG.Vals.I[2,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.151 0.265 0.361

quantile(LMG.Vals.I[3,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.0871 0.1527 0.2121

quantile(LMG.Vals.I[4,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.00759 0.01680 0.07210

quantile(LMG.Vals.I[5,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.00692 0.02045 0.07085

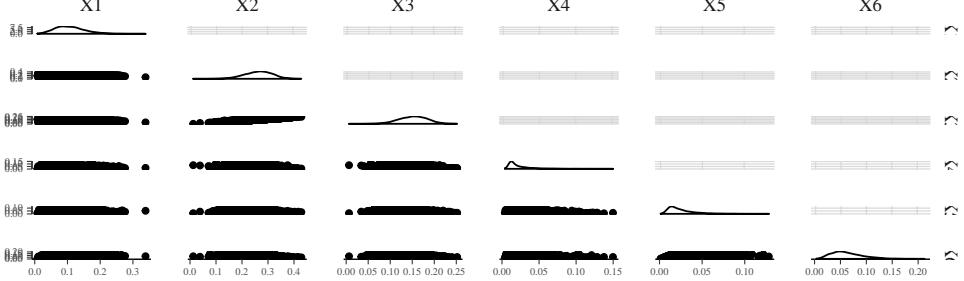
quantile(LMG.Vals.I[6,], c(0.025, 0.5, 0.975))

##    2.5%    50%   97.5%
## 0.0180 0.0577 0.1318

```

```
#Visualization
dat <- data.frame(t(LMG.Vals.I))

pairs.chart <- ggpairs(dat, lower = list(list(combo = "facetdensity")), upper = list(list(conti
pairs.chart


#use prior knowledge
my_prior <- normal(location = c(1, 0.05, -0.03, -3, 0, 0), scale = c(0.1, 0.01, 0.01, 0.1, 0.1, 0.1)

bayes.usair <- stan_glm(SO2 ~ . ,
                         data = usair, prior = my_prior,
                         chains = 4, cores = 4)

#posterior sample
post.sample <- as.matrix(bayes.usair)

#example of the first 10 posterior samples
post.sample[1:10,]

##          parameters
## iterations (Intercept) negtemp manuf      pop   wind  precip    days sigma
## [1,]        106.3   1.213 0.0487 -0.0305 -3.00  0.2553  0.0546 14.8
## [2,]         97.0   0.804 0.0560 -0.0334 -2.95 -0.1191  0.0282 15.3
## [3,]         89.5   0.855 0.0583 -0.0304 -2.96 -0.0258  0.0663 13.8
## [4,]         98.3   0.850 0.0586 -0.0372 -3.05 -0.0598  0.0412 12.7
## [5,]         94.3   0.865 0.0648 -0.0318 -3.04 -0.0657  0.0409 13.5
## [6,]         78.6   0.937 0.0528 -0.0255 -2.97  0.2033  0.1282 13.3
## [7,]        111.7   0.999 0.0557 -0.0268 -3.04 -0.0527 -0.0426 16.1
## [8,]         85.6   0.960 0.0510 -0.0295 -2.95  0.1638  0.1298 12.6
## [9,]         92.9   1.003 0.0645 -0.0373 -3.00  0.0568  0.1140 15.3
## [10,]        101.4   1.118 0.0520 -0.0284 -3.03  0.1328  0.0661 11.8

#no need for the intercept, last parameter is sigma
post.sample <- post.sample[,-1]

#data frame with all submodels
df.rtwos <- rtwos(usair[,2:7], post.sample)

# prepare data frame for LMG values
```

```

LMG.Vals.I<-matrix(0, 6, dim(df.rtwos)[2])

for(i in 1:dim(df.rtwos)[2]){

  gofn<-df.rtwos[,i]

  obj.Gelman<-partition(gofn, pcan = 6, var.names = names(usair[,2:7]))

  LMG.Vals.I[,i]=obj.Gelman$IJ[,1]
}

# posterior LMG distribution of each variable
quantile(LMG.Vals.I[1,], c(0.025, 0.5, 0.975))

##   2.5%    50%  97.5%
## 0.0734 0.1138 0.1594

quantile(LMG.Vals.I[2,], c(0.025, 0.5, 0.975))

##   2.5%    50%  97.5%
## 0.171 0.259 0.334

quantile(LMG.Vals.I[3,], c(0.025, 0.5, 0.975))

##   2.5%    50%  97.5%
## 0.0947 0.1510 0.2083

quantile(LMG.Vals.I[4,], c(0.025, 0.5, 0.975))

##   2.5%    50%  97.5%
## 0.0128 0.0163 0.0214

quantile(LMG.Vals.I[5,], c(0.025, 0.5, 0.975))

##   2.5%    50%  97.5%
## 0.00782 0.01310 0.02503

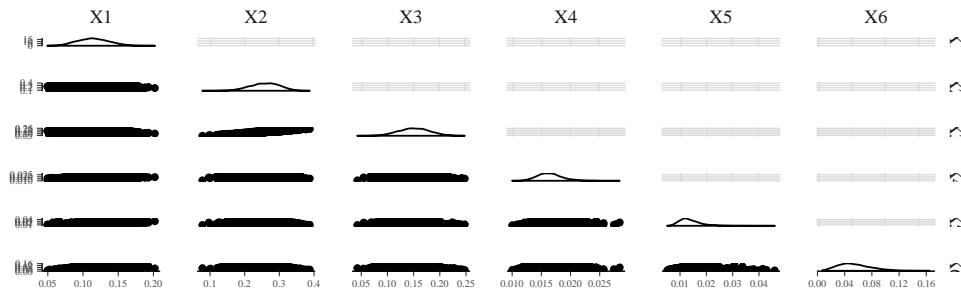
quantile(LMG.Vals.I[6,], c(0.025, 0.5, 0.975))

##   2.5%    50%  97.5%
## 0.0166 0.0515 0.1076

#Visualization
dat <- data.frame(t(LMG.Vals.I))

pairs.chart <- ggpairs(dat, lower = list(list(combo = "facetdensity")), upper = list(list(c
  pairs.chart

```



Chapter 4

Discussion and Outlook - Some extenstions

In the following chapter some extenstions of the LMG formula in the Bayesian framework 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 within subjects can be modeled by including random effects (mixed model) or by assuming correlated errors within a subject (marginal model). Using a random intercept model or a compound symmetry matrix for the error will result in the same model for the fixed predictors. 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.

The extenstion of the LMG formula in the Bayesian framework presented in chapter is restricted to models where the conditional variance formula can easily be applied to get the explained variance of the submodel from the regression parameters of the full model. The focus is on the fixed predictors and not on the random effects. Using the conditional variance formula to get the explained variance of the fixed predictors of the submodels should be applicable in the marginal models, where only the fixed effects are modelled anyway. In the mixed model framework the conditional variance formula is applicable to models including only random intercepts and the focus lies in the explained variance of the fixed predictors. For random-slope models there are atleast some difficulties involved, if it is possible at all the get the expalined variance of the submodel. This chapter shows a random intercept model and a repeated measurement model with an unstructured covariance matrix.

The first example concerns a simple random intercept model with time varying predictors.

4.1 random intercept model

Different R^2 metrics exist for linear mixed models. The variance of a random intercept model with regression parameter β can be written as

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

where $\sigma_f^2 = \text{Var}(\mathbf{X}\beta) = \beta^\top \Sigma_{\mathbf{XX}} \beta$, σ_α^2 is the random intercept and σ_ϵ^2 is the error term.

An R^2 that is guaranteed to be positive can be defined as

$$R_{LMM}^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}, \quad (4.2)$$

Referenz Naka, Snyder.... It is theoretical possible that the R_{LMM}^2 decreases when adding predictors. This may be problematic for the LMG metric, because of violation of the non-negative property. This should rarely be the case with real data. The R^2 can not decrease when adding predictor by using the conditional variance formula on the full model to calculate the R^2 of the submodels. In the Bayesian framework we would draw samples from the posterior distributions of the parameters.

The total variance of the full can be calculated as in equation 6 or by using the samples of the random intercept for each subject directly. The same total variance is then used for one sample of the posterior. (In a repeated measure study we often have within and between subject predictors. If we use the total variance of the full model the random intercept is fitted including all predictors. If a between subject predictor is excluded (e.g. Sex) and we would fit a new random intercept model, the random intercept parameter will in addition explain the variance that was explained by the excluded predictor. In other words it means that the model with the excluded between subject predictor will explain almost as much as the model where the predictor is included when each time a new model and therefore a new random intercept term is fitted in each model.

When using the conditional variance formula for the R^2 of the submodels, it only takes into account the explained variance of the fixed predictors.)

In repeated measurement studies the focus is often in within subject changes. The between subject variance estimated with the random intercept term may not be so important. The more important question may be how much the fixed predictors explain compared to the within subject error, which is

$$R_{\text{repeated}}^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\epsilon^2}, \quad (4.3)$$

The square root of this term is known under the name correlation within subjects by ref(bland Altman 1995). Although in the paper the subject term is fitted as a factor. If we are interested in the within subject effects we can use the model including only the between subject predictor 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 lower importance.

The data are simulated from the following regression setting with $m = 4$ timepoints,

$$Y_{i,j} \sim \mathcal{N}(\beta_0 + x_{1,i,j}\beta_1 + x_{2,i,j}\beta_2 + x_{3,i,j}\beta_3 + x_{4,i,j}\beta_4 + \alpha_i, \sigma^2), \quad i = 1, \dots, n \quad j = 1, \dots, m \quad (4.4)$$

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

The following R code is used to simulate the data:

```
sub<- 1:20
subi <- rnorm(20, 0, 4)
subi<-rep(subi, 4)
t <- c(0, 1, 2,3)
t <- c(rep(0, 20), rep(1,20), rep(2, 20), rep(3,20))
```

```

mu <- rep(0,4)
sig <- matrix(0.4, 4, 4)
diag(sig) <- 1
sig[3,4] <- 0.9
sig[4,3] <- 0.9
sig[1,2] <- 0.3
sig[2,1] <- 0.3

rawvars <- mvrnorm(n=80, mu=mu, Sigma=sig)

x1 <- t+rawvars[,1]
x2 <- t+rawvars[,2]
x3 <- t+rawvars[,3]
x4 <- t+rawvars[,4]

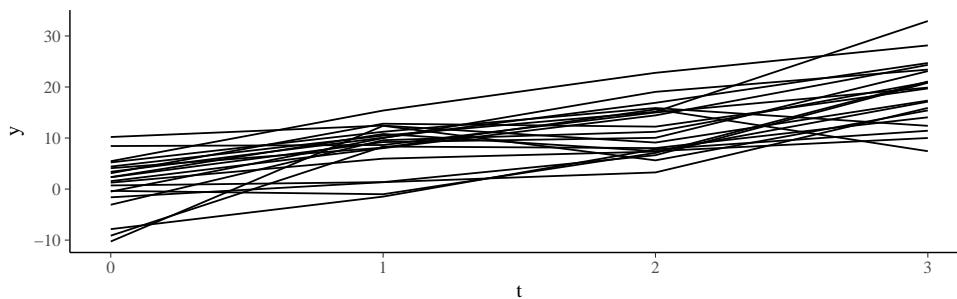
b1 <- b2 <-1
b3 <- b4 <- 2

y<- x1*b1 +x2*b2 +x3*b3+ x4*b4 + subi+ rnorm(80, 0, 0.1)

df <- data.frame(y=y, x1 = x1, x2=x2, x3 = x3, x4 = x4, sub = rep(sub,4))

p <- ggplot(data = df, aes(x = t, y = y, group = sub))
p + geom_line()

```



The R^2 of the submodels were first calculated according to the formula in equation,. The random intercept effect is not of interest. We see that most of the within subject variance is explained by the predictors. The credible intervals are very narrow. For the information about the between subject variance we can look at the random intercept variance directly. In the second part the random intercept is included in the total variance calculation and the calculation of the R^2 values. We see that there is a large between subject variance in this dataset. The LMG values including the between subject variance are much lower. The credible intervals are also much wider, because the uncertainty about the between subject variance is also included. In my opinion we can get more useful information from separating the between and within variance components when it is easy possible. 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-parameteric bootstrap, resampling whole subjects (all repeated measurements of a subject).

```

fit <- stan_glmer(y ~ x1+x2+x3+x4 + (1|sub) ,
                     data = df,

```

```

chains = 4, cores = 4

post.sample <- as.matrix(fit)
post.sample.r <- post.sample[,c(2:5,(ncol(post.sample)-1):ncol(post.sample))]

df.rtwos <- rtwos.ri.r(df[,2:5], post.sample.r)

LMG.Vals<-matrix(0, 4, dim(df.rtwos)[2])

for(i in 1:dim(df.rtwos)[2]){

  gofn<-df.rtwos[,i]

  obj.Gelman<-partition(gofn, pcan = 4, var.names = names(df[,2:5]))

  LMG.Vals[,i]=obj.Gelman$IJ[,1]
}

# posterior LMG distribution of each variable
quantile(LMG.Vals[1,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.204 0.205 0.207

quantile(LMG.Vals[2,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.207 0.208 0.210

quantile(LMG.Vals[3,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.292 0.293 0.295

quantile(LMG.Vals[4,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.291 0.293 0.294

# explained compared to total variance

df.rtwos <- rtwos.ri.a(df[,2:5], post.sample)

LMG.Vals<-matrix(0, 4, dim(df.rtwos)[2])

for(i in 1:dim(df.rtwos)[2]){

  gofn<-df.rtwos[,i]

  obj.Gelman<-partition(gofn, pcan = 4, var.names = names(df[,2:5]))
}

```

```

LMG.Vals[,i] = obj.Gelman$IJ[,1]
}

# posterior LMG distribution of each variable
quantile(LMG.Vals[1,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.035 0.090 0.158

quantile(LMG.Vals[2,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.0489 0.0992 0.1444

quantile(LMG.Vals[3,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.0623 0.1222 0.1755

quantile(LMG.Vals[4,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.0666 0.1296 0.1857

```

4.2 marginal model

The next example concerns a repeated measurement model with an unstructured covariance error structure. The data are generated from the following model:

$$Y_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}), \quad i = 1, \dots, n \quad (4.5)$$

where $\boldsymbol{\Sigma}$ represents an unstructured error covariance matrix.

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. When there are no missing timepoints in each subject, we can take the mean of $\text{diag}(\boldsymbol{\Sigma})$ to make the formula compatible with the $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \mathbf{X} \boldsymbol{\beta}$ of equation resulting in the total variance term

$$\text{Var}(\mathbf{Y}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \mathbf{X} \mathbf{X}^\top \boldsymbol{\beta} + \text{mean}(\text{diag}(\boldsymbol{\Sigma})), \quad (4.6)$$

The following R-code is used to generate the data:

```

sub<- 1:20
subi <- rnorm(20, 0, 1)
subi<-rep(subi, 4)

mu <- rep(0,4)
sig <- matrix(0.4, 4, 4)
diag(sig) <- 1

```

```

sig[3,4] <- 0.9
sig[4,3] <- 0.9
sig[1,2] <- 0.3
sig[2,1] <- 0.3

rawvars <- mvrnorm(n=80, mu=mu, Sigma=sig)
cov(rawvars)

##      [,1]  [,2]  [,3]  [,4]
## [1,] 0.942 0.292 0.400 0.401
## [2,] 0.292 0.866 0.455 0.425
## [3,] 0.400 0.455 1.402 1.176
## [4,] 0.401 0.425 1.176 1.184

t <- c(rep(1, 20),rep(2,20), rep(3, 20), rep(4, 20))
x1 <- t+rawvars[,1]
x2 <- t+rawvars[,2]
x3 <- t+rawvars[,3]
x4 <- t+rawvars[,4]

Sig<- matrix(3, 4,4)
diag(Sig) <- 10
u <- rep(0, 4)
Sig[1,1] <- 5
Sig[2,2] <- 7
Sig[3,4] <- 8
Sig[4,3] <- 8

Sig[1,2] <- 4
Sig[2,1] <-4

Sig

##      [,1]  [,2]  [,3]  [,4]
## [1,]     5     4     3     3
## [2,]     4     7     3     3
## [3,]     3     3    10     8
## [4,]     3     3     8    10

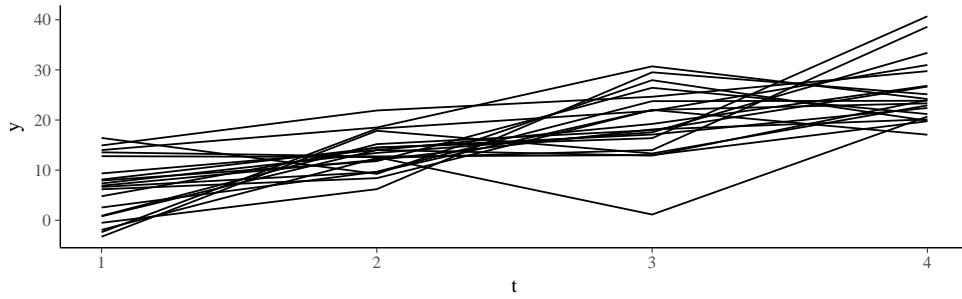
error <- mvrnorm(20, u, Sig)

y<- x1*b1 +x2*b2 + x3*b3 +x4*b4 +c(error)

t <- c(rep(1, 20), rep(2, 20), rep(3, 20), rep(4, 20))
df <- data.frame(y=y, x1 = x1, x2=x2 , x3 = x3, x4 = x4, sub = rep(sub,4), t =t)

p <- ggplot(data = df, aes(x = t, y = y, group = sub))
p + geom_line()

```



```
# Bayesian framework
```

```
Y <- matrix(df[, 'y'], 20, 4, byrow=F)
x1 <- matrix(df[, 'x1'], 20, 4, byrow=F)
x2 <- matrix(df[, 'x2'], 20, 4, byrow=F)
x3 <- matrix(df[, 'x3'], 20, 4, byrow=F)
x4 <- matrix(df[, 'x4'], 20, 4, byrow=F)
```

```
N = 20 #subjects
M = 4 # repeated measures
```

```
#-----
```

```
modelString <- "model{

# Likelihood
for(i in 1:N){
Y[i,1:M] ~ dmvnorm(mu[i,1:M], Omega[1:M,1:M])
for(j in 1:M){
mu[i,j] <- beta0 + beta1*x1[i,j]+ beta2*x2[i,j]+ beta3*x3[i,j] + beta4*x4[i,j]
} }

# Priors

Omega[1:M, 1:M] ~ dwish(zRmat[1:M,1:M] , zRscal)
Sigma[1:M, 1:M] <- inverse(Omega)

beta0      ~ dnorm(0,0.001)
beta1      ~ dnorm(0,0.001)
beta2      ~ dnorm(0,0.001)
beta3      ~ dnorm(0,0.001)
beta4      ~ dnorm(0,0.001)

}"
```

```
writeLines( modelString , con="Jags-MultivariateNormal-model.txt" )
```

```
model <- jags.model(textConnection(modelString),
```

```
data = list(Y=Y)
```

```

n.chains=3)

## Compiling model graph
## Resolving undeclared variables
## Allocating nodes
## Graph information:
##   Observed stochastic nodes: 20
##   Unobserved stochastic nodes: 6
##   Total graph size: 789
##
## Initializing model

samp <- coda.samples(model,
variable.n
n.ITER=200

summary(samp)

##
## Iterations = 1:20000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 20000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean      SD Naive SE Time-series SE
## Sigma[1,1] 5.807 2.107 0.008602      0.00955
## Sigma[2,1] 3.863 1.777 0.007254      0.00800
## Sigma[3,1] 2.775 2.306 0.009416      0.00980
## Sigma[4,1] 4.562 3.058 0.012483      0.01426
## Sigma[1,2] 3.863 1.777 0.007254      0.00800
## Sigma[2,2] 5.077 2.307 0.009418      0.01104
## Sigma[3,2] 2.912 2.776 0.011332      0.01241
## Sigma[4,2] 5.027 3.853 0.015731      0.01662
## Sigma[1,3] 2.775 2.306 0.009416      0.00980
## Sigma[2,3] 2.912 2.776 0.011332      0.01241
## Sigma[3,3] 11.218 5.084 0.020756      0.02137
## Sigma[4,3] 12.660 6.143 0.025080      0.02548
## Sigma[1,4] 4.562 3.058 0.012483      0.01426
## Sigma[2,4] 5.027 3.853 0.015731      0.01662
## Sigma[3,4] 12.660 6.143 0.025080      0.02548
## Sigma[4,4] 18.737 8.630 0.035230      0.03585
## beta1       1.153 0.194 0.000790      0.00155
## beta2       0.812 0.175 0.000716      0.00125
## beta3       2.438 0.440 0.001797      0.01002
## beta4       1.792 0.476 0.001941      0.01114
##
## 2. Quantiles for each variable:
##
##           2.5%     25%     50%     75%  97.5%

```

```

## Sigma[1,1] 3.047 4.384 5.40 6.760 10.92
## Sigma[2,1] 1.553 2.720 3.56 4.655 7.95
## Sigma[3,1] -0.587 1.440 2.51 3.809 7.63
## Sigma[4,1] 0.185 2.709 4.16 5.948 11.19
## Sigma[1,2] 1.553 2.720 3.56 4.655 7.95
## Sigma[2,2] 2.583 3.786 4.71 5.937 9.62
## Sigma[3,2] -0.250 1.597 2.63 3.912 7.59
## Sigma[4,2] 1.009 3.202 4.59 6.329 11.53
## Sigma[1,3] -0.587 1.440 2.51 3.809 7.63
## Sigma[2,3] -0.250 1.597 2.63 3.912 7.59
## Sigma[3,3] 5.930 8.491 10.45 13.038 20.72
## Sigma[4,3] 6.232 9.390 11.74 14.852 24.20
## Sigma[1,4] 0.185 2.709 4.16 5.948 11.19
## Sigma[2,4] 1.009 3.202 4.59 6.329 11.53
## Sigma[3,4] 6.232 9.390 11.74 14.852 24.20
## Sigma[4,4] 9.950 14.214 17.47 21.746 34.73
## beta1      0.769 1.025 1.15 1.282 1.53
## beta2      0.469 0.696 0.81 0.926 1.16
## beta3      1.562 2.147 2.44 2.733 3.29
## beta4      0.876 1.472 1.79 2.107 2.73

#LMG calculations

samp <- coda.samples(model,
variable.names
n.iter=20000,

post.sample <- samp[[1]][,c(5:8, 1:4)]

df.rtwos <- rtwos.marg(df[,2:5], post.sample, 4) # 4 repeated measures

LMG.Vals<-matrix(0, 4, dim(df.rtwos)[2])

for(i in 1:dim(df.rtwos)[2]){

  gofn<-df.rtwos[,i]

  obj.Gelman<-partition(gofn, pcan = 4, var.names = names(df[,2:5]))

  LMG.Vals[,i]=obj.Gelman$IJ[,1]
}

# posterior LMG distribution of each variable
quantile(LMG.Vals[1,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.100 0.128 0.150

quantile(LMG.Vals[2,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.0987 0.1267 0.1462

```

```
quantile(LMG.Vals[3,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.166 0.213 0.245

quantile(LMG.Vals[4,], c(0.025, 0.5, 0.975))

## 2.5% 50% 97.5%
## 0.167 0.213 0.240
```

Chapter 5

Other variable importance metrics in the Bayesian framework

Different variable importance metrics exist. The conditional variance formula allowed us to calculate the R^2 of the submodels from the posterior sample of the full model. The focus of this master thesis was on the LMG formula. For each posterior sample the LMG formula can be applied for the submodels of each posterior sample. A lot of the variable importance metrics are based on the R^2 of the full model compared to the submodels. Instead of the LMG formula we could as well have used another variable importance metric after we have calculated the R^2 of all the submodels. Commonality and dominance analysis seem to be interesting extensions of the LMG framework. Both provide besides the LMG information some other information about the variance decomposition of the predictors. The relaimpo package provides some more bootstrap options like pariwise differences that could also be easily transferred to the Bayesian framework.

5.1 Conclusion

The Bayesian frameworks provides us with a nice option to include the uncertainty about parameters. Posterior distributions of the parameters allow us to calculate a distribution of R^2 values for each model. Using the conditional variance formula allowd us 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. Doing it this way has some nice properties. The interdependence of the submodels to each other is respected. The R^2 of the submodels does not decrease when adding predictors. The non-negativity of the shares is therefore respected in the LMG formula. This property of using the conditional variance formula is also interesting when the LMG formula is applied to random intercept models fitted by maximum likelihood.

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.

Assuming non-stochastic or stochastic predictors can have a big impact on the the uncertainty of the explained variance and the LMG values. Although the posterior regression parameter distribtuions is the same in both cases (under some assumptions described in chapter ...) the explainened variance of a model is directly dependent on the covarinace 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.

Chapter 6

Appendix

Bibliography

- Alexander, D. L. J., Tropsha, A., and Winkler, D. A. (2015). Beware of R(2): Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of chemical information and modeling*, **55**, 1316–22. [5](#)
- Gelman, A., Goodrich, B., Gabry, J., and Ali, I. (2017). R squared for Bayesian regression models *. Technical report. [5](#)
- Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *American Statistician*, **61**, 139–147. [1](#)
- Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**, 137–152. [1](#), [4](#)
- Kvalseth, T. O. (1985). Cautionary Note about R 2. *The American Statistician*, **39**, 279. [5](#)
- 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](#)

