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 ref groemping 2007) in linear models estimated with Bayesian methods.

Regression models are popular in many applied research areas. These models provide a tool to find an association between a response variable Y and a set of explanatory variables. These 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. Goering 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 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. 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. A summary of the metrics can be found in ref (groemping, 2015). Groemping

2015 recommends to use the existing best practices, until a more profound solution is found. For variance (or genereally goodness of fit) decomposition based importance she recommend to use LMG enhanced with joint contributions or dominance analysis.

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. The following formulas are taken from Groemping 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.

$$\text{evar}(S) = \text{Var}(Y) - \text{Var}(Y \mid X_j, j \in S), \tag{1.1}$$

$$\text{svar}(M \mid S) = \text{evar}(M \cup S) - \text{evar}(S), \tag{1.2}$$

, where S and M denote disjoint sets of predictors.

$R^2(S)$ can be written as $\text{evar}(S)/\text{Var}(Y)$.

The LMG formula is given below for the first predictor only. Because of exchangeable predictors, this is no loss of generality.

$$\begin{aligned}
\text{LMG}(1) &= \frac{1}{p!} \sum_{\pi permutation} \text{svar}(\{1\} \mid S_1(\pi)), \\
&= \frac{1}{p!} \sum_{S \subseteq \{2,\ldots,p\}} n(S)!(p - n(S) - 1)!\text{svar}(\{1\} \mid S) \\
&= \frac{1}{p} \sum_{i=0}^{p-1} \left(\sum_{\substack{S \subseteq \{2,\ldots,p\} \\ n(S)=1}} \text{svar}(\{1\} \mid S) \right) \Bigg/ \binom{p-1}{i} \qquad (1.3) \\
&= \frac{1}{p} \sum_{i=0}^{p-1} \frac{\sum_{\substack{S \subseteq \{2,\ldots,p\} \\ n(S)=1}} \text{svar}(\{1\} \mid S)}{\binom{p-1}{i}} \qquad (1.4) \\
& \qquad (1.5)
\end{aligned}$$

, 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 sequentiel 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.

The LMG metric is implemented in the R packages relaimpo and hier.part.

Chevan and Sutherland propose that instead of the variances an appropriate goodness-of-fit metric can 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. Hierarchical partitioning is implemented in the hier.part package. 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).

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:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.6}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.7}$$

, where $\hat{y}_i == \mathrm{E}(y \mid X_n, \hat{\theta})$.

When other estimation methods than maximum likelihood are used (4) can be negative and (5) can be bigger than 1. When the model is fitted by Baysian methods and estimates of the posteriors are employed this can easily happen (Gelman). A model that explaines 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. When new data from a test set is predicted (e.g. in leave-one-out crossvalidation), one has to be careful with the R^2 definition used. R^2 should then be calculated by equation(4). Negative R^2 can then also happen, when the model is fitted by maximum likelihood (Referenz Beware of R2). Because of R^2 bigger than one or the possibility of negative R^2 values Gelman(2018) proposes to use:

$$R^2_{Gelman} = \frac{V(\sum_{n=1}^{N} \hat{y}_n^s)}{V(\sum_{n=1}^{N} \hat{y}_n^s) + V(\sum_{n=1}^{N} e_n^s)} \tag{1.8}$$

, where $\hat{y}_n^s = E\left(y \mid X_n, \theta^s\right)$ and the vector of errors $e_n^s = y_n - \hat{y}_n^s$, V stands vor the variance defined as $V(\sum_{n=1}^{N} z_n) = \sum_{n=1}^{N}(z_n - \bar{z})^2/(N-1)$ for any vector z and $\theta^s, s = 1, ..., S$ are draws from the posterior parameter distribution. The formula is then guaranteed to be between 0 and 1. However, they argue, that 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. Especially for the LMG formula this may be problematic. The R^2 can then 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.

The variance of the linear model can be written as

$$\mathrm{Var}(y) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta} + \sigma^2, \tag{1.9}$$

where

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

Writting it this way makes it clear that the predictors in the Gelman equation can also be taken as random (Gelman 2017).

Another possible R^2 for the classical linear model would be to sample the σ parameter from the posterior distribution instead of defining it as in equation (6). This would then yield to the following R^2 definition:

$$R^2_{Snyder} = \frac{V(\sum_{n=1}^{N} \hat{y}_n^s)}{V(\sum_{n=1}^{N} \hat{y}_n^s) + \sigma^s} \tag{1.10}$$

, where $\hat{y}_n^s = E\left(y \mid X_n, \theta^s\right)$, V stands vor the variance and $\theta^s, s = 1, ..., S$ are draws from the posterior parameter distribution.

As a side note: Instead of using $V(\sum_{n=1}^{N} e_n^s)$ we could also use $\sum(y - \hat{y}^s)^2/(n-1)$ as an estimate for the error. For the maximum likelihood estimate $var(y_i - \hat{y_i}) = \sum(y_i - \hat{y_i})^2/(n-1)$. 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. $var(y_i - \hat{y_i}) = \sum(y_i - \hat{y_i})^2/(n-1)$ is than a little bit bigger than $var(y_i - \hat{y_i})$. 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. When we use $var(y_i - \hat{y_i})$. 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. However, when we think of the Bayes R^2 as an estimate of goodness-of-fit for new data negative values may not be something that should be strictly avoided. As in the case of predicting new data from a test set, equation (4) can reasonably be applied in the Bayes case, leading to:

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y - \hat{y}_n^s)^2}{\sum_{n=1}^{N}(y - \bar{y})^2}, \tag{1.11}$$

, where $\hat{y}_n^s = E\left(y \mid X_n, \theta^s\right)$ and $\theta^s, s = 1, ..., S$ are draws from the posterior paramter distribution.

For two predictors equation 6 simplifies to

$$\mathrm{Var}(y) = \beta_1^2 \mathrm{Var}(X_1) + 2\beta_1\beta_2 \mathrm{Cov}(X_1, X_2) + \beta_2^2 \mathrm{Var}(X_2) + \sigma^2, \tag{1.12}$$

A benefit of writting the variance in this way, is that it is in the linear model possible to calculate the explained variance of the submodels from the fullmodel including all predictors.

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 equation 7 . In mathematical notation that is

$$\operatorname{svar}(X_1 \mid \emptyset) = \beta_1^2 \operatorname{Var}(X_1) + 2\beta_1\beta_2 \operatorname{Cov}(X_1, X_2) + \beta_2^2 \operatorname{Var}(X_2)\rho_{12}^2 \tag{1.13}$$

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 can be used to calculate the R^2 of all submodels.

As an examle the conditional distribution of a normal distribution: 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

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

The conditional variance formula can be used to calculate the R^2 of the submodels. Lets assume we are interested in the R^2 of a model containining the predictors $\mathbf{X}_{q\ldots p}$, and 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 normal distribution example above we have the covariance matrix of p predictors written as

$$\boldsymbol{\Sigma}_{\mathbf{XX}} = \operatorname{Cov}(\mathbf{X}) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{p\times p}, \tag{1.14}$$

$$\boldsymbol{\Sigma}_{11} = \operatorname{Cov}(\mathbf{X}_{1\ldots q-1}, \mathbf{X}_{1\ldots q-1}), \tag{1.15}$$

$$\boldsymbol{\Sigma}_{12} = \operatorname{Cov}(\mathbf{X}_{1\ldots q-1}, \mathbf{X}_{q\ldots p}), \tag{1.16}$$

$$\boldsymbol{\Sigma}_{22} = \operatorname{Cov}(\mathbf{X}_{q\ldots p}, \mathbf{X}_{q\ldots p})$$

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

$$\operatorname{Cov}(\mathbf{X}_{1\ldots q-1} \mid \boldsymbol{x}_{q\ldots p}) = \boldsymbol{\Sigma}_{11} \ - \ \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{1.17}$$

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

$$\text{evar}(\mathbf{X}_{1...p}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}. \tag{1.18}$$

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

$$\text{evar}(\mathbf{X}_{q...p}) = \text{evar}(\mathbf{X}_{1...p}) - \boldsymbol{\beta}_{1...q-1}^\top \text{Cov}(\mathbf{X}_{1...q-1} \mid \boldsymbol{x}_{q...p}) \boldsymbol{\beta}_{1...q-1}. \tag{1.19}$$

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

$$\text{evar}(\mathbf{X}_{q...p}) / \text{Var}(\mathbf{Y}). \tag{1.20}$$

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 paramter 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.

1.1 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 $\boldsymbol{y}$ on $\mathbf{X}$. The conditional mean of a continuous response variable $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$) is related to a $n \times k$ predictor matrix $\mathbf{X}$ via a linear model,

$$\mathrm{E}(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}, \tag{1.21}$$

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

$$\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \tag{1.22}$$

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

$$p(\boldsymbol{y}, \mathbf{X} \mid \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\theta}_{y|x})\, p(\mathbf{X} \mid \boldsymbol{\theta}_x), \tag{1.23}$$

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 in the posterior parameters $\boldsymbol{\theta}_{y|x}$. These posterior densities are proportional to likelihood of the data multiplied by the prior densitiy. The joint density $p(\boldsymbol{y}, \mathbf{X} \mid \boldsymbol{\theta})$ is used to learn about the posterior parameters, via Bayes Rule

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathbf{X}) \propto p(\boldsymbol{y}, \mathbf{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{1.24}$$

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

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}_x \mid \boldsymbol{y}, \mathbf{X}) = \frac{p(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y} \mid \mathbf{X})} \times \frac{p(\mathbf{X} \mid \boldsymbol{\theta}_x)\, p(\boldsymbol{\theta}_x)}{p(\mathbf{X})} \tag{1.25}$$

This factorization shows that under the above mentioned assumptions the posterior inference about the parameters $\boldsymbol{\theta}_{y|x} = (\beta, \sigma^2)$ is indepdendent 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 $\theta_{y|x} = (\boldsymbol{\beta}, \sigma^2)$ 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 caclulations 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.