Chapter 1

Theoretical background

1.1 LMG variable importance metric

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

The following notations for the explained variance (1.1) and sequentially added variance (1.2) simplify the notation of the LMG formula.

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

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

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

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

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

The different formula writings help to better understand what the calculation is about in the LMG metric. The R^2 of the model including all predictors is decomposed. In the formula on the top (1.3), the LMG value of predictor 1 is represented as an unweighted average over all orderings of the sequential added variance contribution of predictor 1. The formula in the center (1.4), shows that the calculation can be done more efficiently. The orderings with the same set of predecessors S are combined into one summand. Instead of $p!$ summands only 2^{p-1} summands need to be calculated. The formula on the bottom (1.5) shows that the LMG metric can also be seen as the unweighted average over average explained variance improvements when adding predictor 1 to a model of size i without predictor 1 (Grömping, 2015). The LMG metric is implemented in the R package `relaimpo` (Grömping, 2006).

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

1.2 Appropriate R^2 definitions in the Bayesian framework

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

Two commonly used R^2 definitions are:

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

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

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

When other estimation methods than maximum likelihood are used, equation (1.6) can be negative and equation (1.7) 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 is nonsense. A negative R^2 is also difficult to interpret. A negative R^2 may be interpreted as a fit that is worse than the mean of the data. This can make sense for predictive purposes, e.g. when new data from a test set is predicted by leave-one-out crossvalidation (Alexander *et al.*, 2015). For non predicting purposes, a negative R^2 does not make sense. The aim of the LMG formula is to gain some more information about the possible association between variables. A predictor can not explain less than zero variance in the population. To respect the non-negative share property of the LMG formula, the R^2 of submodels should not decrease when adding predictors. Both classical R^2 definitions seem not to be well suited for the LMG metric in the Bayesian framework.

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

$$\mathrm{Var}(y) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta}) + \sigma^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta} + \sigma^2, \tag{1.8}$$

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 predictors.

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

$$\mathrm{R}^2_{Gelman} = \frac{\mathrm{Var}(\sum_{i=1}^n \hat{y}_i^s)}{\mathrm{Var}(\sum_{i=1}^n \hat{y}_i^s) + \mathrm{Var}(\sum_{i=1}^n e_i^s)}, \qquad i = 1, \dots, n, \tag{1.9}$$

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

In the Bayesian framework, the σ^2 parameter is explicitly modeled in the standard linear regression setting. Therefore, it is possible to sample the σ^2 parameter from its posterior distribution instead of defining the error as in definition (1.9), which would lead to the following definition:

$$\begin{aligned} \mathrm{R}^2 &= \frac{\mathrm{Var}(\sum_{i=1}^n \hat{y}_i^s)}{\mathrm{Var}(\sum_{i=1}^n \hat{y}_i^s) + \sigma_s^2} \\ &= \frac{\boldsymbol{\beta}_s^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}_s}{\boldsymbol{\beta}_s^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}_s + \sigma_s^2}, \qquad i = 1, \dots, n, \end{aligned} \tag{1.10}$$

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

The predictors in definition (1.9) and definition (1.10) could also be taken as random (Gelman *et al.*, 2017). The predictors are then called stochastic predictors. Using the sample covariance estimate provides then just an estimate of the true covariance structure. With stochastic predictors, there is an additional uncertainty in the R^2 formula that can have a large influence on the R^2 and the LMG values.

In practice, definition (1.10) and definition (1.9) should lead to similar values in the standard linear model. In my opinion, it is more reasonable to take the full Bayesian route by sampling σ^2 of its posterior distribution. This approach provides the opportunity to include prior information about σ^2 directly into to R^2 calculations. The LMG calculations in the examples of this master thesis will therefore be based on definition (1.10). A benefit of definition (1.9) is that it also works for generalized linear models where we often have no separate variance parameter.

The denominator of R^2 is no longer fixed in definition (1.9) and in definition (1.10). We can therefore no longer interpret an increase in R^2 as an improved fit to a fixed target (Gelman *et al.*, 2017). The unfixed denominator seems to be problematic for the LMG formula in the Bayesian framework. However, in the linear model it is possible to calculate the R^2 of all submodels from the parameters of the model inlcuding all predictors (full-model) and the covariance matrix of the predictors. Therefore, all submodels of a posterior sample are compared to the same fixed value. A possible way to get the R^2 of the submodels from the full-model is shown in the next section.

1.3 Use of conditional variance formula to get R^2 of submodels

For two predictors, definition (1.8) simplifies to

$$\operatorname{Var}(y) = \beta_1^2 \operatorname{Var}(X_1) + 2\beta_1\beta_2 \operatorname{Cov}(X_1, X_2) + \beta_2^2 \operatorname{Var}(X_2) + \sigma^2, \tag{1.11}$$

When predictor X_1 is the only one in the model, the explained variance includes the variance of the predictor itself, the whole covariance term, and some of the contribution of the variance of X_2 in equation (1.11) additionaly . In mathematical notation, that is

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

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

In the general case with p regressors, the conditional variance formula (1.12) can be used to calculate the R^2 of all submodels. For example, the conditional variance formula can be used to specify the conditional distribution of a multivariate normal distribution $\mathbf{Y}$.

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

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

The joint distribution is a multivariate normal distribution with elements

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

the conditional distribution is normally distributed again with mean

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

and the conditional variance is

$$\operatorname{Var}(\mathbf{Y}_1|\mathbf{y}_2) = \boldsymbol{\Sigma}_{11} \; - \; \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{1.12}$$

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

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

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

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

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

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

The total explained variance of the full-model containing $\mathbf{X}_{1\dots p}$ omits simply the σ^2 parameter in (1.8) , which is

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

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

$$\operatorname{evar}(\mathbf{X}_{q\dots p}) = \operatorname{evar}(\mathbf{X}_{1,\dots,p}) - \boldsymbol{\beta}_{1,\dots,q-1}^\top \operatorname{Cov}(\mathbf{X}_{1,\dots,q-1} \mid \boldsymbol{x}_{q\dots p}) \boldsymbol{\beta}_{1,\dots,q-1}. \tag{1.13}$$

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

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

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

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

Instead of using the conditional mean formula to get the R^2 of the submodels, it would be possible to fit a separate Bayesian model for each submodel. An R^2 distribution can easily be built for each submodel by using definition (1.9) or definition (1.10). However, the problem is how to calculate the LMG values out of these R^2 distributions. If we just sample independently from the R^2 distributions, the dependence of the paramter values of the submodels to each other is ignored. We would have many possibly true parameter values of a predictor in the same LMG comparison. It would then also be possible that the R^2 decreases when adding predictors. Another drawback is that it would be much more time-consuming to fit a separate Bayesian model for each submodel. Using the conditional variance formula on the full-model allows to calculate LMG values in the Bayesian framework in a reasonable time exposure. Depending on the number of predictors and the number of posterior samples, the calculations still take some time in the Bayesian framework. For stochastic predictors, the computation time is multiplied by the number of covariance samples.

1.4 Bayesian Regression

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

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

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

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

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

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

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

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

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

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

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

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

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}_x \mid \boldsymbol{y}, \mathbf{X}) = \frac{p(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y} \mid \mathbf{X})} \times \frac{p(\mathbf{X} \mid \boldsymbol{\theta}_x)\, p(\boldsymbol{\theta}_x)}{p(\mathbf{X})}. \tag{1.14}$$

The factorization in equation 1.14 shows, that under the above mentioned assumptions, the posterior inference about the parameters $\boldsymbol{\theta}_{y|x} = (\beta, \sigma^2)$ is independent from the inference about $\boldsymbol{\theta}_x$ given data $\mathbf{X}$. This also means that the assumptions about $\mathbf{X}$ being non-stochastic or stochastic result in the same posterior density of $\boldsymbol{\theta}_{y|x}$. In the case of non-stochastic regressors, $p(\mathbf{X})$ and $\boldsymbol{\theta}_x$ drop out of the calculations. For stochastic predictors, it means, that given $\mathbf{X}$, nothing more can be gained about $\theta_{y|x} = (\boldsymbol{\beta}, \sigma^2)$ from knowing $\boldsymbol{\theta}_x$.

The focus of regression is on $\boldsymbol{\theta}_{y|x} = (\boldsymbol{\beta}, \sigma^2)$, for which it does not matter whether we assume fixed or stochastic predictors under the above mentioned assumptions. The variance of the predictors is also incorporated in the LMG formula. The LMG formula may be especially interesting for continuous predictors, which often are of stochastic nature. Grömping (2006) recommends in most cases to use the non fixed regressor option when calculating bootstrap confidence intervals. Therefore, the information about $\boldsymbol{\theta}_x$ would also be relevant for stochastic regressors. As seen in equation (1.14), inference about $\boldsymbol{\theta}_x$ is independent from inference about $\boldsymbol{\theta}_{y|x}$. If there are stochastic predictors and we use the sample estimate of the covariance matrix, we do not incorporate the uncertainty of the estimate. Because the explained variance is calculated by $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}$, inference about $\boldsymbol{\theta}_x$ seems to be equally important as inference about $\boldsymbol{\theta}_{y|x}$ for stochastic predictors. If the distribution of the $p(\mathbf{X})$ is known, the $\boldsymbol{\theta}_x$ could be estimated. However, the computation time is then much higher, because the whole LMG calculation need to done for each posterior covariance sample of the predictors. Depending on the number of predictors this would be very time-consuming. In most cases, the problem is that the distribution of the $\mathbf{X}$ is unknown. As a practical solution, nonparametric bootstrapping of the covariance matrix could be used to include the uncertainty of the stochastic predictors in the LMG calculations. Again, it would be necessary to do the LMG calculations for each bootstrap sample of the covariance matrix. There exist also different covariance estimators. The shrinkage method may be an interesting estimator with some nice properties (Schäfer and Strimmer, 2005).

Bibliography

Alexander, D. L., Tropsha, A., and Winkler, D. A. (2015). Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling*, **55**, 1316–1322. 2

Chevan, A. and Sutherland, M. (1991). Hierarchical partitioning. *American Statistician*, **45**, 90–96. 2

Gelman, A., Goodrich, B., Gabry, J., and Ali, I. (2017). R-squared for Bayesian regression models *. Technical report. 3

Grömping, U. (2006). Relative Importance for Linear Regression in R : The Package relaimpo. *Journal of Statistical Software*, **17**, 1–27. 2, 7

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *American Statistician*, **61**, 139–147. 4

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**, 137–152. 1, 2

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley. 6

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

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, 1–30. 7

Walsh, C. and Nally, R. M. (2015). Title Hierarchical Partitioning. Technical report. 2