

# Chapter 1

## Methods

### 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 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.

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

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

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

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

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

The different writings of the formula help to better understand 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` (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 predictors itself is included in the model, compared to its independent component (I) is named the joint contribution (J). Hierarchical partitioning is implemented in the `hier.part` package (Walsh and Nally, 2015). The LMG values are calculated when  $R^2$  is chosen as the goodness-of-fit measure. The partitioning function of the `hier.part` package is used in this master thesis. The hierarchical partitioning function accepts as input a data frame with the  $R^2$  values of all submodels. Of note, the `hier.part` package is only guaranteed to work for 9 predictors and does not work at all for more than 12 predictors.

## 1.2 Appropriate $R^2$ definitions in the Bayesian framework

The focus of this master thesis is on the linear model. 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:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1.3)$$

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad i = 1, \dots, n, \quad (1.4)$$

, 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 (1.3) can be negative and (1.4) 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.

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

$$\text{Var}(y) = \text{Var}(\mathbf{X}\boldsymbol{\beta}) + \sigma^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{X}} \boldsymbol{\beta} + \sigma^2, \quad (1.5)$$

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

By using this variance definition [Gelman \*et al.\* \(2017\)](#) proposes to use

$$R_{Gelman}^2 = \frac{\text{Var}(\sum_{i=1}^n \hat{y}_i^s)}{\text{Var}(\sum_{i=1}^n \hat{y}_i^s) + \text{Var}(\sum_{i=1}^n e_i^s)}, \quad i = 1, \dots, n, \quad (1.6)$$

, 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](#)).

In the Bayesian framework the  $\sigma^2$  parameter is explicitly modeled in the standard linear regression setting. It is therefor possible to sample the  $\sigma^2$  parameter from its posterior distribution instead of defining the error as in (1.6), which would lead to the following definition:

$$R^2 = \frac{\text{Var}(\sum_{i=1}^n \hat{y}_i^s)}{\text{Var}(\sum_{i=1}^n \hat{y}_i^s) + \sigma_s^2} \quad (1.7)$$

$$= \frac{\boldsymbol{\beta}_s^\top \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{X}} \boldsymbol{\beta}_s}{\boldsymbol{\beta}_s^\top \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{X}} \boldsymbol{\beta}_s + \sigma_s^2} \quad (1.8)$$

, 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.

The predictors in (1.6) and (1.8) could also be taken as random ([Gelman \*et al.\*, 2017](#)). The predictors are then called stochastic predictors. Using the sample covariance estimate provides us 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 especially the LMG values.

In practice (1.8) and (1.6) 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  $\sigma^2$ . This provides us the opportunity to include prior information about  $\sigma^2$  directly into to  $R^2$  calculations. The LMG calculations in the examples of this master thesis will therefore be based on (1.8). The benefit of (1.6) 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 (1.6) and (1.8). We can therefore no longer interpret an increase in  $R^2$  as an improved fit to a fixed target (Gelman *et al.*, 2017). 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 full model 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.

### 1.3 Conditional variance formula

For two predictors (1.5) simplifies to

$$\text{Var}(y) = \beta_1^2 \text{Var}(X_1) + 2\beta_1\beta_2 \text{Cov}(X_1, X_2) + \beta_2^2 \text{Var}(X_2) + \sigma^2, \quad (1.9)$$

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 (1.9). 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 (Grömping, 2007).

In the general case with  $p$  regressors, the conditional variance formula (1.10) 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

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

and the conditional variance is

$$\text{Var}(\mathbf{Y}_1 | \mathbf{y}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \quad (1.10)$$

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

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

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

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

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

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

The total explained variance of the full model  $\mathbf{X}_{1\dots p}$  omits simply the  $\sigma^2$  parameter in (1.5), which is

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

The explained variance of a submodel can be calculated by subtracting from the total explained variance the explained 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  $\mathbf{X}_{q,\dots,p}$  can therefore be written as

$$\text{evar}(\mathbf{X}_{q,\dots,p}) = \text{evar}(\mathbf{X}_{1,\dots,p}) - \boldsymbol{\beta}_{1,\dots,q-1}^\top \text{Cov}(\mathbf{X}_{1,\dots,q-1} \mid \mathbf{x}_{q,\dots,p}) \boldsymbol{\beta}_{1,\dots,q-1}. \quad (1.11)$$

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

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

where  $\text{Var}(\mathbf{Y})$  is defined as  $\boldsymbol{\beta}^\top \Sigma_{\mathbf{X}\mathbf{X}} \boldsymbol{\beta} + \sigma^2$ .

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 parameters 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 property that all shares should be non-negative and the dependence of the parameter values 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 (1.6) or (1.8). 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 parameter 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 us to calculate LMG values in the Bayesian framework in a reasonable amount of time. 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

This Section provides a short 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  $\mathbf{y}$  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,

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

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

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

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

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

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

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

The dependence of  $\mathbf{y}$  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

$$p(\beta, \sigma^2, \boldsymbol{\theta}_x \mid \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2) p(\beta, \sigma^2)}{p(\mathbf{y} \mid \mathbf{X})} \times \frac{p(\mathbf{X} \mid \boldsymbol{\theta}_x) p(\boldsymbol{\theta}_x)}{p(\mathbf{X})}. \quad (1.12)$$

This factorization 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  $\boldsymbol{\theta}_{y|x} = (\beta, \sigma^2)$  from knowing  $\boldsymbol{\theta}_x$ .

The focus of regression is on  $\boldsymbol{\theta}_{y|x} = (\beta, \sigma^2)$ , for which it does not matter whether we assume fixed or stochastic predictors under the above mentioned assumptions. For the LMG formula, the variance of the predictors is also incorporated. The LMG formula may be especially interesting for continuous predictors, which often are of stochastic nature. For stochastic predictors the information about  $\boldsymbol{\theta}_x$  would therefore also be relevant. As seen in (1.12) 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  $\beta^\top \Sigma_{\mathbf{X}\mathbf{X}} \beta$ , 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 nice properties (Schäfer and Strimmer, 2005).





# Bibliography

- Alexander, D. L., Tropsha, A., and Winkler, D. A. (2015). Beware of R<sup>2</sup>: 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](#), [4](#)
- Grömping, U. (2006). Relative Importance for Linear Regression in R : The Package relaimpo. *Journal of Statistical Software*, **17**, 1–27. [2](#)
- 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<sup>2</sup>. *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](#)