Chapter 1

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 1.1 everything is properly scaled.

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.

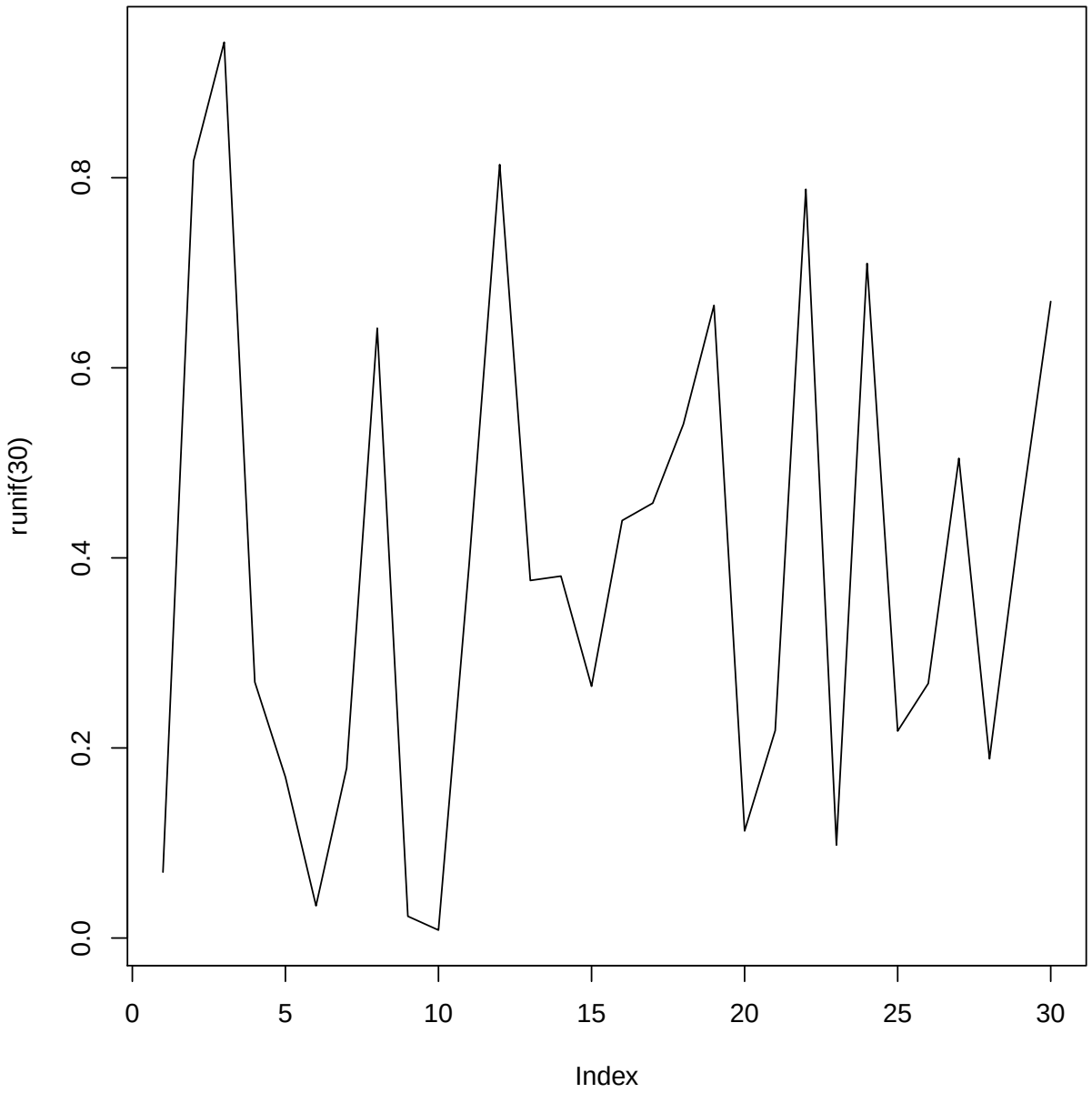

Figure 1.1: Test figure to illustrate figure options used by knitr.

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

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

, where S and M denote disjoint sets of predictors.

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

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

, 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 (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:

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

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

, 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 (1.6) can be negative and (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 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. 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:

$$\mathrm{R}^2_{Gelman} = \frac{\mathrm{Var}(\sum_{i=1}^{n} \hat{y}_i^s)}{\mathrm{Var}(\sum_{i=1}^{n} \hat{y}_i^s) + \mathrm{Var}(\sum_{i=1}^{n} e_i^s)}, \qquad i = 1, \ldots, n, \tag{1.8}$$

, 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 $\theta^s, s = 1, ..., S$ 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

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

where $\boldsymbol{\beta}^\top = (\beta_1 \ldots \beta_p)$ 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 (1.8) 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 (1.8), which woud lead to the following definition:

$$R^2 = \frac{\mathrm{Var}(\sum_{i=1}^{n} \hat{y}_i^s)}{\mathrm{Var}(\sum_{i=1}^{n} \hat{y}_i^s) + \sigma^s} \tag{1.10}$$

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

In practice (1.10) and (1.8) should lead to similar values in the standard linear model. In my opionion it is more reasonable to go the full bayesian route and sample σ^2. This provides us the opportunity to include prior information abouth σ^2 directly into to R^2 calculations. The LMG calculations in the examples of this master thesis will therefore be based on (1.10). The benefit of (1.8) is that it also works for generalized linear models, where we often have no separate variance parameter.

For two predictors (1.9) 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 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 (1.11) . 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.

In the general case with p regressors, the conditional variance formula (1.12) 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

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

and the conditional variance is

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

We want to calculate R^2 of a submodel 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

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

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

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

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

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

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

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

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

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

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

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,

$$\mathrm{E}(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}, \tag{1.14}$$

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

$$\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \tag{1.15}$$

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

$$p(\boldsymbol{y}, \mathbf{X} \mid \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\theta}_{y|x})\, p(\mathbf{X} \mid \boldsymbol{\theta}_x), \tag{1.16}$$

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

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathbf{X}) \propto p(\boldsymbol{y}, \mathbf{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{1.17}$$

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

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}_x \mid \boldsymbol{y}, \mathbf{X}) = \frac{p(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y} \mid \mathbf{X})} \times \frac{p(\mathbf{X} \mid \boldsymbol{\theta}_x)\, p(\boldsymbol{\theta}_x)}{p(\mathbf{X})} \tag{1.18}$$

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.

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. 4

Gelman, A., Goodrich, B., Gabry, J., and Ali, I. (2017). R squared for Bayesian regression models *. Technical report. 4

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

Kvalseth, T. O. (1985). Cautionary Note about R 2. *The American Statistician*, **39**, 279. 3