Variable importance in regression models

Ulrike Grömping

Beuth University of Applied Sciences Berlin

This is the peer reviewed version of the following article:
Grömping, U. (2015). Variable importance in regression models. *WIREs Comput Stat* **7**, 137-152.***, which has been published in final form at https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1346. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.***

Abstract

Regression analysis is one of the most-used statistical methods. Often part of the research question is the identification of the most important regressors or an importance ranking of the regressors. Most regression models are not specifically suited for answering the variable importance question, so that many different proposals have been made. This article reviews in detail the various variable importance metrics for the linear model, particularly emphasizing variance decomposition metrics. All linear model metrics are illustrated by an example analysis. For non-linear parametric models, several principles from linear models have been adapted, and machine-learning methods have their own set of variable importance methods. These are also briefly covered. Although there are many variable importance metrics, there is still no convincing theoretical basis for them, and they all have a heuristic touch. Nevertheless, some metrics are considered useful for a crude assessment in the absence of a good subject matter theory.

INTRODUCTION

Regression analysis in its general form – investigating the influence of a set of regressors (*X*-variables) on a response variable of interest (the *Y*-variable) – is one of the most important topics in applied statistics. There are many types of regression models, differing in the types of response variable that can be investigated and in the strength of parametric assumptions made. Regardless of the type of model, applied researchers often request an assessment of the relative importance of the different regressors for the response variable. Most types of regression models are not designed for directly answering this question. Intuitively, many researchers first think of the sizes of (standardized) coefficients or test statistics for determining variable importance; however, such simple metrics have substantial limitations for assessing variable importance. Consequently, the assessment of variable importance requires additional considerations and has been the subject of controversial discussions. Also, since it has been addressed from researchers in many different fields, some metrics have been reinvented various times (e.g. Fabbris[1], Genizi[2] and Johnson[3] all proposed the same metric), which does shed lide on them from different angles.

The paper is written in terms of the univariate regression situation, i.e. a single response variable (left-hand side variable) *Y* is modeled depending on a set of explanatory variables, which are also called independent variables, predictor variables, or – throughout the paper – regressors. Throughout the parametric part of the paper, it is assumed that each model effect has one df only, and that all effects are on the same level of hierarchy (i.e., the model does not include interactions which would have to be excluded whenever any of their parent effects are excluded); for some variable importance metrics – particularly game-theory-based variance decomposition metrics – it is easily possible to remove this

assumption, i.e., to incorporate explanatory variables with more than one degree of freedom or interaction effects, which introduce a hierarchy among the effects. The implementation is reasonably straightforward; these situations are not covered, however, in order to avoid additional notational complexity.

The next section covers simple methods for linear regression models, namely raw regression coefficients, t-tests, raw correlations, standardized regression coefficients, semipartial correlations, a simple method proposed originally by Hoffman[4] and advocated by Pratt[5], and sequential decomposition of R^2. The key controversies regarding variable importance assessment are already present for the simple metrics, and are also discussed in the next section. The section "General concepts and fields of application" gives an overview of types of model, types of question, and fields of application for variable importance assessment. A discussion of common requirements for variable importance metrics for the – best-researched – linear model sheds light on common ground and conflicts regarding the nature of variable importance. Subsequently, the most detailed section "Methods based on variance decomposition" discusses the most widely used advanced metrics for variable importance in linear regression, and the subsequent section exemplifies all variable importance methods for the linear model. The shorter sections "Variable importance for parametric non-linear models" and "Variable importance in machine learning methods" give an overview of further methods, mainly focusing on the hierarchical partitioning approach by Chevan and Sutherland[6] and information-based variable importance for the parametrics case and on random forest variable importance for machine learning. All sections provide hints on where the methods are implemented in the open source R statistical software[7]. The final section discusses the state of the field and future directions.

SIMPLE METRICS FOR MEASURING RELATIVE IMPORTANCE IN LINEAR REGRESSION MODELS

In this section, we assume a linear regression model of the form

$$Y_i = \beta_0 + X_{i1}\beta_1 + \ldots + X_{ip}\beta_p + \varepsilon_i \qquad (1)$$

with independent error terms ε_i of expectation 0 and constant positive variance σ^2. The marginal correlation between Y_i and X_{ij} is denoted as ρ_j, the semipartial (or part) correlation as $\rho_{j\cdot\text{other}}$, and the variances of the X_{ij} as σ_j^2; the estimates for β, ρ and σ parameters are denoted by the letters b, r and s, respectively. The following simple metrics are frequently discussed for ranking the variables X_j in terms of importance for Y:

(i) the absolute values or squares of the raw coefficients b_j;
(ii) the absolute values or squares of t-values t_j or the p-values p_j from t-tests of the null hypotheses $\beta_j=0$, $j=1,\ldots,p$;
(iii) the absolute values or squares of the raw correlations r_j; the squared version is identical to R^2 values from univariate models;
(iv) the absolute values or squares of the *semipartial* correlations $r_{j\cdot\text{other}}$ (also called *part* correlations); the squared version is identical to the reduction in R^2 when removing the regressor X_j from the full model with all p regressors, or – in other words – the squared correlation of Y with the residuals from regressing X_j on all other variables (what does X_j contribute to Y over and above the other regressors);
(v) the absolute values or squares of the standardized coefficients $b_{j,st} = b_j\, s_j / s_y$;
(vi) the products $b_{j,st}\, r_j$, as proposed by Hoffman[4] and advocated by Pratt[5],

(vii) the sequential increase in R^2 (equivalent to the sequential increase in the model sum of squares, known as Type I SS), when entering each regressor to the model in a pre-specified order.

All these metrics can be obtained from standard regression and/or correlation output. If squares are chosen in metrics (iii) to (v), the metrics (iii) to (vii) sum to the multiple R^2 from the regression model in case of p uncorrelated regressors. For correlated regressors, this remains true only for metric (vi) – which is one of the main reasons why Hoffman proposed it – and for metric (vii). Many users of relative importance methods in linear models find it desirable to be able to decompose R^2; however, they typically expect decomposition into non-negative portions, whereas metric (vi) can and does yield negative contributions in relevant cases. Therefore, metric (vi) has been rejected by many authors – including Ward[8], Darlington[9], Bring[10] and the present author. Pratt[5] attempted a justification of metric (vi) based on a set of axioms that are satisfied by this metric alone; Thomas, Hughes and Zumbo[11] also advocated the method. Today's advocates of metric (vi) agree that the cases with one or more negative products should be treated as abnormal, and relative importance conclusions should not be drawn without deep-diving the root cause of the negative shares. Metric (vii) is the other metric that is able to decompose R^2 in case of correlated regressors; however, this metric is likewise unacceptable because of its dependence on the order of regressors, except for very rare cases for which the regressors may come with a natural order. Note that the built-in R function **anova** delivers this order-dependent sequential decomposition of the model variance. For correlated regressors, it is thus usually preferable to use the function **Anova** from R package **car** (Fox and Weisberg[12]), which allocates to each regressor a variance share in the spirit of metric (iv); this share is not order-dependent, but consequently does not decompose the overall model sum of squares, except for the case of uncorrelated regressors.

Metrics (i) to (v) have also been controversially discussed. Of course, as the raw coefficients b_j and the t-values t_j are most easily available, the reason for suggesting different metrics must be dissatisfaction with their properties. The raw coefficient b_i reflects the influence on the response of a unit change in the ith regressor, given fixed values of the other regressors. The most important disadvantage of the raw coefficients is that they are not scale invariant: for example, the coefficient of a regressor "height" can be dramatically increased by changing the unit from cm to m. On the other hand, Achen[13] – who devoted Chapter 6 of his book on "interpreting and using regression" to "the importance of a variable" – considered the raw beta coefficients as the appropriate metric for assessing theoretical importance if there is a natural scale. Darlington[9] presented a similar line of reasoning for the standardized coefficient. Bring[10], on the other hand, used geometric reasoning to criticize the use of standardized coefficients for situations with correlated variables. The present author sides with Bring[10], as it seems that the division by the regressor variance in the standardized coefficient introduces an artificial element that is neither appropriate when asking the question of "theoretical importance" as considered in Achen[13] and nor adequately reflects the influence on response variability. Metric (ii) (ranking according to t-values or p-values) is equivalent to metric (iv). These were discussed by both Darlington[9] (as metric (iv), called "usefulness") and Bring[10] (as the t-test). Both these authors acknowledged the merits of these metrics in particular for prediction. Darlington[9] also discussed metric (iii), the raw correlation (called "validity"), and acknowledged its benefit particularly in its ignorance about which other variables are in the model. t-values, p-values and semipartial correlations consider each variable conditional on all other variables in the model, while raw correlations, on the contrary, consider each variable on its own (marginal perspective). Budescu[14] and Johnson and Lebreton[15] emphasized that reasonable metrics for assessing variable importance have to incorporate both perspectives: conditional and marginal. These two antipoles occur again for more complex metrics, e.g. in the comparison between the computer-intensive variance decomposition metrics

PMVD and LMG (see Section “Methods based on variance decomposition”) or conditional and CART-based random forests (see Section “Variable importance in machine learning methods”).

GENERAL CONCEPTS AND FIELDS OF APPLICATION

Diversity of concepts for variable importance

Besides the above-mentioned theoretical importance – also called causal importance by other authors, Achen[13] proposed two further types of importance in the context of linear models:

- level importance – in a given situation for the current mean of all regressors, what is the average influence on the level of the response – answered by $b_j \cdot \text{mean}(X_j)$
- and dispersion importance – for which he proposed the standardized coefficient $b_{j,\text{st}}$ as the appropriate answer and simultaneously criticized the omnipresence of this answer in spite of the fact that he considers dispersion importance (in this sense) as not usually of interest.

A recent example for an application of level importance can be found in Holgersson, Norman and Tavassoli[16], who seem to have rediscovered the method; strangely, they justified their preference of this method over the more common variance decomposition methods by its simplicity –in the author’s point of view, the decision between level importance and variance decomposition should be taken based on the nature of the research question.

Achen’s[13] understanding of dispersion importance is different from that of many other authors who consider variance decomposition methods as the appropriate tool for assessing dispersion importance. Several metrics for variance decomposition in the linear model will be investigated more closely in the next section – among them the game-theory-based metrics LMG and PMVD, which are based on averaging metric (vii) over all orderings of regressors.

Outside the linear model, there are also proposals for the assessment of variable importance. This article covers the following approaches: Extending Theil and Chung[17], Chevan and Sutherland[6] proposed to assess variable importance in the generalized linear model by hierarchical partitioning, which generalizes the principle of averaging over orderings of regressors. Machine-learning methods have special ways of assessing variable importance; this article specifically discusses random forest variable importance.

Outside the scope of this article, there are several related approaches: Silber, Rosenbaum and Ross[18] presented an approach for comparing the relative importance of two sets of variables in general regression models; their approach was implemented by Firth[19] in the R package **relimp**. For log-linear Poisson models, Ortmann[20] took an approach similar to Achen’s[13] level importance and suggested a computer-intensive game-theoretic approach for calculating it; Land and Gefeller[21] already proposed a similar approach for risk partitioning in epidemiology. Other authors partitioned risk differently, e.g. Eide and Gefeller[22]. A further approach to variable importance comes from a causal inference perspective (van der Laan[23], Ritter et al.[24]). Gevrey, Dimopoulos and Lek[25] discussed methods for variable importance assessment in neural networks; their work is instrumental for the neural network implementation of the variable importance function **varImp** of R package **caret** by Kuhn[26]. All these latter approaches are not covered in this article. Furthermore, many authors use variable importance measures for variable selection. Contrary to this perspective, this article considers the variable importance for a given set of variables and a given model only.

All variable importance approaches discussed in this article relate to a (relatively) simple overall regression model that refrains from detailing causal structures; if one is willing/able to work with more

detailed models like structural equation models, these rather coarse variable importance approaches are no longer suitable. This is already emphasized by Pedhazur[27] (Section 7 of the monograph).

Requirements for relative importance metrics

Several authors have formulated requirements for relative importance metrics, some of which are widely agreed in the scientific community, while others are controversial. A somewhat arbitrary collection of common criteria is described and discussed below, and Table 1 details which of the metrics for assessing variable importance in the linear model satisfies which of the requirements; the table includes the simple metrics from the previous section as well as the variance decomposition metrics from the following section.

(a) *Anonymity*: *the relative importance is not affected by the labels / positions of the regressors.* Game theorists explicitly voice this need (e.g. Ortmann[28]). Although the criterion is seldom mentioned in statistical literature, it is usually taken for granted. Note, however, that the simple metric (vii) (sequential variance decomposition) violates this criterion in case of correlated regressors.

(b) *Relative importance does not depend on anything but the first two moments of the joint distribution of the variables*.
This requirement is voiced e.g. by Pratt[5] (axiom A1). Like the previous one, it is implicitly respected by all methods for the linear model, because linear models and all relevant quantities can be estimated from the mean vector and the variance-covariance matrix (plus the information on the number of observations for the degrees of freedom of sampling distributions). The only exception is the simple metric (vii), because it additionally depends on the order of regressors.

(c) *Relative importance is not changed by linear transformations on individual variables* (e.g. Pratt[5], axiom A2). This scale invariance requirement is respected by most metrics – only the raw beta coefficient does not respect this requirement.

(d) *The addition of a pure noise variable, independent of y and $x_1 \ldots x_p$, to a subset of variables does not affect the importance of the subset relative to the other variables*.
This requirement – also brought forward by Pratt[5] (axiom A5) – is again implicitly respected by all methods.

(e) *Relative importance should balance out conditional and marginal considerations.* This requirement was brought forward by Budescu[14] and later also by Johnson and Lebreton[15], who implied that the contribution of X_j when alone in the model (called direct effect), the contribution of X_j in addition to all other regressors (called total effect) and the contributions of X_j considering different subsets of further regressors should all be reflected in an appropriate relative importance metric.

(f) *Proper decomposition: the model variance is decomposed, that is, the sum of all shares is the model variance (or R^2, depending on normalization).*
This is the defining requirement for variance decomposition metrics that will be discussed in the next section.

(g) *Orthogonal compatibility: The decomposition respects orthogonal subgroups, i.e. for each orthogonal subgroup of regressors, the assigned shares sum to the unique overall model variance (or R^2) for that subgroup.*
This requirement for variance decomposition metrics was explicitly stated by Genizi[2]. All the variance decomposition metrics satisfy it.

(h) *Non-negativity: all allocated shares are always non-negative.*
This requirement for variance decomposition metrics is backed by many authors, as is apparent from the discussions regarding the Pratt decomposition, which has been almost

ridiculed because of its failure to always yield non-negative shares. All other variance decomposition metrics discussed in this article satisfy this requirement.

(i) *Exclusion: the share allocated to a regressor X_j with $\beta_j = 0$ should be 0.*
This requirement was introduced by Feldman[29], who proposed the PMVD decomposition which is the only known method to satisfy both requirements (h) and (i). The Pratt decomposition and the simple metrics that involve estimated coefficients also satisfy this requirement.
This requirement is particularly convincing for advocates of a conditional approach to variable importance: a variable should be allocated an importance of zero, if it has zero impact, given all other variables. This aspect is discussed in detail later.

(j) *Inclusion: a regressor X_j with $\beta_j \neq 0$ should receive a nonzero share.*
This requirement is satisfied by all variance decomposition metrics, except for the Pratt metric for which it could be violated in constellations for which the marginal correlation is zero in spite of a non-zero coefficient. The simple squared correlation can also violate this requirement under special circumstances.

(k) *If m+n equicorrelated variables with equal impact on the response y are combined into two sum variables x_1 (sum of m variables) and x_2 (sum of n variables), the relative importance of these two should be like m to n.*
This requirement – introduced by Pratt[5] (axiom A3) – is exclusively satisfied by the Pratt decomposition. No other metrics attempt to satisfy it.

(l) *The nonsingular linear transformation of a subset ($x_1 \ldots x_q$) into the subset (x'_1, x'_q) does not affect its importance relative to the other variables.*
Again, the requirement – introduced by Pratt[5] (axiom A6) – is exclusively satisfied by the Pratt decomposition. No other metrics attempt to satisfy it.

The last two requirements are satisfied by the Pratt decomposition only. In the author's view, requirement (k) becomes plausible only, if the correlation between regressors is zero. In that case, all decompositions fulfill it. Otherwise, the covariances of X_1 and X_2 with Y still have the relation m/n; however, the covariance between X_1 and X_2 is non-negligible and does – of course – have an impact on the allocation of shares. Requirement (l) sounds reasonable at first glance; however, a non-singular linear transformation for a group of variables that is not an orthogonal subgroup does not only affect the group of variables itself but also its covariance structure with the other variables and thus should not reasonably be expected to deliver an unchanged overall share of that group of variables.

TABLE 1 ABOUT HERE

Fields of application

In 1989, Kruskal and Majors[35] reported a literature survey in which they sampled articles from many different fields, e.g. chemistry, economy, banking, social science and medicine, genetics, psychology, legal studies, political science, history. This list covers many but not all of the fields where there is still active use and development of methods for variable importance. Ecology sees a particularly widespread use of methods for variable importance, e.g. MacNally[36,37], MacNally and Walsh[38], or Gevrey, Dimopoulos and Lek[25]. Management and organizational research is another field with substantial attention to variable importance, e.g. Soofi, Retzer and Yasai-Ardekani[39], Johnson and Lebreton[15], Lebreton, Ployhart and Ladd[40] or Nimon and Oswald[41]. Recently, relative importance of variables has also received substantial attention in sensory analysis (Bi and Chung[42], Bi[43]). Closely

related, market research has long been an area where variable importance analysis is extensively studied – often under the heading "(key) driver analysis". For example, Lipovetsky and Conklin[44] proposed the so-called "Shapley value regression" which decomposes R^2 into LMG shares (see next section) and modifies coefficient estimates accordingly (the adjustment of coefficients was fundamentally criticized by Grömping and Landau[45]). Note that market research uses two different applications of the game-theoretic Shapley value, one of which is about partitioning risk / market share rather than the classical variable importance in regression. Epidemiologists also use methods for partitioning risk among several potential risk factors (called exposures in their context), rather than the classical variable importance methods (see also the section "Diversity of concepts for variable importance").

METHODS BASED ON VARIANCE DECOMPOSITION

This section discusses six methods that decompose the model variance or – equivalently – the coefficient of determination R^2 in linear regression analysis: The first part of this section presents the computer-intensive methods LMG (Lindeman, Merenda and Gold[32], Kruskal[33]), dominance analysis (Budescu[14]) as an extension thereof, and PMVD (Feldman[29]) as a modification for satisfying the exclusion requirement. Subsequently, the less computationally demanding decompositions by Gibson[30], Green, Carroll and DeSabro[46], Fabbris[1], Genizi[2], Johnson[3] and Zuber and Strimmer[31] are discussed. As these six proposals contain several reinventions of the same approach, they add only three further decompositions to the portfolio. The Hoffman/Pratt method, which was already presented as simple metric (vi), is excluded from the discussion of variance decomposition methods, because it can and does yield negative shares in practically relevant situations. Likewise, the sequential variance decomposition (simple metric (vii)) is excluded because of its order dependence.

The model variance of model (1) can be written as

$$\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{XX}\boldsymbol{\beta} = \sum_{j=1}^{p}\sum_{k=1}^{p}\beta_j\beta_k\sigma_{jk}\,, \tag{2}$$

where

- $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1 \ldots \beta_p)$ does not include the intercept parameter,
- $\boldsymbol{\Sigma}_{XX}$ is the true unknown p x p covariance matrix of the regressors with elements σ_{jk} and can be written as $diag_j\left(\sqrt{\sigma_{jj}}\right)\mathbf{P}_{XX}\,diag_j\left(\sqrt{\sigma_{jj}}\right)$
- with $\mathbf{P}_{XX}$ (capital Rho) the theoretical correlation matrix.

Summed with the error variance σ^2, (2) yields the total variance var(Y); R^2 is the proportion of (2) in the total variance. In the following, for simplicity, the data for model (1) are assumed to be centered, i.e. the response vector $\mathbf{Y}$ and the nxp predictor matrix $\mathbf{X}$ have column means 0, so that the empirical covariance matrices can be written as $\mathbf{S}_{XX} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/(n-1)$ and $\mathbf{S}_{XY} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}/(n-1)$, respectively. As this centering does not affect R^2 or the model variance, this simplification does not reduce generality of the considerations. The estimated model variance $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{S}_{XX}\hat{\boldsymbol{\beta}}$ is consistent for

$$\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}{}^{-1}\boldsymbol{\Sigma}_{XY} = \sigma^2\mathbf{P}_{YX}\mathbf{P}_{XX}{}^{-1}\mathbf{P}_{XY}\,, \tag{3}$$

with $\mathbf{P}_{YX}$ and $\mathbf{P}_{XY}$ the unknown correlation vectors (row or columns, respectively) of the response Y with the regressors. In presenting the simple metrics, the elements of the empirical analogs $\mathbf{b}$, $\mathbf{S}_{XX}$, $\mathbf{R}_{XX}$,

$\mathbf{R}_{YX}$ and s_y^2 of these theoretical quantities have been used; as the empirical analogs are consistent estimators for the theoretical quantities, use of these two sets of symbols is exchangeable.

Variance decomposition in linear models with correlated regressors is still a topic for research and discussion, because (2) has a natural unique decomposition into summands for uncorrelated regressors only (i.e. if $\sigma_{jk}=0$ for $j\neq k$). Whenever predictors are correlated, the customary sequential variance decomposition (Type I SS, simple metric (vii)) depends on the order of the variables, whereas the customary unique "each variable last" variance allocation (Type II or Type III SS, equivalent to simple metric (iv)) yields allocations that do not sum to the overall model variance.

LMG, dominance analysis and PMVD:
computer-intensive methods related to game theory

There is a common agreement that proper decomposition into non-negative shares is required, and that order dependence is usually unacceptable. LMG and PMVD were introduced as an unweighted (LMG: Lindeman, Merenda and Gold[32]; Kruskal[33,34]) or weighted (PMVD: Feldman[29]) average of sequential explained variances over all possible orderings of regressors. Dominance analysis (Budescu[14]; Azen and Budescu[47], Budescu and Azen[48]) is a more detailed variant of LMG and is presented in the second part of this section.

LMG and PMVD

In the following formulae, evar and svar denote the explained variance and sequentially added variance, respectively:

$$\text{evar}(S) = \text{var}(Y) - \text{var}(Y|X_j,\, j \in S) \tag{4}$$

and $$\text{svar}(M\,|S) = \text{evar}(M\cup S) - \text{evar}(S), \tag{5}$$

where S and M denote disjoint sets of predictors.

LMG and PMVD can be directly written in terms of these expressions. For notational simplicity, their formulae are given below for the first predictor – as predictor labels are exchangeable, this is no loss of generality. With $S_1(\pi)$ the set of predecessors of predictor 1 in permutation π, three representations of LMG are useful:

$$\begin{aligned}
\text{LMG}(1) &= \frac{1}{p!}\sum_{\pi\ permutation} \text{svar}(\{1\}\,|\,S_1(\pi)) \\
&= \frac{1}{p!}\sum_{S\subseteq\{2,\ldots,p\}} n(S)!(p-n(S)-1)!\text{svar}(\{1\}\,|\,S) \\
&= \frac{1}{p}\sum_{i=0}^{p-1}\left(\sum_{\substack{S\subseteq\{2,\ldots,p\},\\ n(S)=i}} \text{svar}(\{1\}\,|\,S)\Big/\binom{p-1}{i}\right)
\end{aligned} \tag{6}$$

The top formula represents LMG as an unweighted average over all orderings of the sequential contribution of predictor 1, the middle formula is computationally more efficient, since it combines orderings with the same set of predecessors S into one summand, and the bottom formula by Christensen[49] shows that LMG is the unweighted average over average contributions to models of different sizes.

PMVD is also an average over orderings of the sequential contributions of predictor 1, however a weighted one:

$$\mathrm{PMVD}(1) = \sum_{\pi \text{ permutation}} p(\pi)\,\mathrm{svar}(\{1\} \mid S_1(\pi)). \tag{7}$$

The weights are $p(\pi) = L(\pi)/\sum_\pi L(\pi)$, with

$$L(\pi) = \prod_{i=1}^{p-1} \mathrm{svar}\left(\{\pi_{i+1},...,\pi_p\} \mid \{\pi_1,...,\pi_i\}\right)^{-1} = \prod_{i=1}^{p-1} \left(\mathrm{evar}\left(\{1,...,p\}\right) - \mathrm{evar}\left(\{\pi_1,...,\pi_i\}\right)\right)^{-1}. \tag{8}$$

These weights strongly favor orderings π for which the early factors have a large contribution. They have been constructed such that the exclusion requirement holds, i.e., a regressor with zero coefficient gets the weight zero (Feldman[29]). Note that computation of PMVD should be based on a different representation, which makes use of the underlying game-theoretic proportional value and its potential; however, even exploiting this efficiency improvement, the computational burden for PMVD is higher than that for LMG because the number of summands cannot be reduced by considering subsets rather than orderings.

Both LMG and PMVD are related to game theory: Stufken[50] explained the connection of LMG to the game-theoretic Shapley value (Shapley[51]), whereas Feldman[52] worked out the connection of PMVD to the proportional value (Ortmann[28]). In both cases, the regressors are considered as the players in a cooperative game, the explained variance achievable by a set of regressors (a coalition) is the worth attached to that coalition, and the overall explained variance achievable by all regressors together as the worth of the grand coalition or the total gain that is to be distributed fairly among all players=regressors. The Shapley value and the proportional value operate on two different sets of axioms. These also lead to two different sets of properties for LMG and PMVD; however, so far nobody has worked out the consequences of these from a statistical point of view. Both metrics decompose R^2 or – equivalently – the explained variance into non-negative shares. PMVD has been designed to fulfill the exclusion criterion, while LMG does not fulfill that criterion. From a conditional perspective, exclusion is a natural requirement: a coefficient 0 in the model with all regressors indicates an importance of zero, given all other regressors in the model. However, Grömping[53] argued that exclusion is not necessarily a useful requirement in case of correlated regressors, since under some causal structures the coefficient 0 for the regressor does by no means indicate an unimportant variable. Therefore, under ignorance about causal structures, exclusion is not a reasonable requirement at least if the importance question has been asked from a theoretical / causal point of view. Even though both LMG and PMVD incorporate all variable orderings, LMG is closer to the marginal perspective, while PMVD is closer to the conditional one, so that the conflict between these two perspectives remains relevant.

LMG and PMVD are implemented in the R-package **relaimpo** (Grömping[54]); the offer of PMVD is restricted to non-US users, however, since PMVD is patented in the US.

Dominance analysis

Dominance analysis allocates the LMG overall share to each regressor. However, it does not stop there: in addition, the difference to the univariate R^2 of each regressor is called the "joint contribution" of the variable. Besides allocating these two overall metrics for each regressor, dominance analysis compares each pair of variables, distinguishing between complete dominance and general dominance: variable A generally dominates variable B, if its LMG allocated variance (see formula (6) below) is larger than that of variable B. Even if variable A generally dominates variable B, for a certain set S of other variables in the model, the additional contribution of variable B may be larger than that of variable A. If variable A completely dominates variable B, A has a larger contribution than B regardless which other variables are present in the model. While the concept of complete dominance is

interesting, it makes the pattern of situations to be looked at quite complex and confusing; as dominance analysis remains a coarse method that does not model e.g. a causal structure, it might be better to invest effort into refining the model rather than into putting more detail into an inherently crude analysis of relative importance. The author therefore does not advocate the most detailed form of dominance analysis. However, the joint contributions (see also Chevan and Sutherland[6] for generalized linear models) are a useful supplement to the LMG variance decomposition. Note that – contrary to the allocated variance shares – some of the joint contributions may become negative. Commonality analysis (see e.g. Ray-Mukherjee et al.[55] for a recent account of this old method) is somewhat similar to dominance analysis, but does not even attempt to obtain p shares for p regressors; rather, it decomposes R^2 into 2^p–1 shares; for the same reason for which the more detailed aspects of dominance analysis are considered of limited use, commonality analysis is not further detailed here. The R-package **yhat** (Nimon and Oswald[41]) provides dominance analysis as well as commonality analysis; dominance analysis can also be obtained as a special case of R-package **hier.part** (Walsh and MacNally[56]; safe to use for up to 9 regressors only). Azen[57] provided SAS macros for dominance analysis. In his PhD thesis, Fickel[58] (in German) developed another deep-dive of the contributions to the overall variable importance of a regressor. Unfortunately, his notation was overly complicated, which presumably prevented his work from being published. An unpublished English language version of his interesting contribution is available from a preprint server[59].

Gibson decomposition / CAR scores, Green et al. decomposition and Fabbris / Genizi / Johnson decomposition

In this section, additionally to the already established column centering, columns of the **X** matrix and the **Y** vector are standardized, i.e. have empirical variance 1. Again, this reduction does not affect R^2 or any relative assessments of model sums of squares, so that the simplification does not reduce generality of the considerations. For these normalized data, the empirical correlation matrices are $\mathbf{R}_{XX} = \mathbf{X}^T\mathbf{X}/(n-1)$ and $\mathbf{R}_{XY} = \mathbf{X}^T\mathbf{Y}/(n-1)$. Like discussed in connection with formulae (2) and (3), empirical quantities and their theoretical counterparts can be used interchangeably, having in mind that the empirical expressions are consistent estimators for the theoretical quantities.

Gibson[30], Green, Carroll and DeSarbo[46], Fabbris[1], Genizi[2], Johnson[3] and Zuber and Strimmer[31] all introduced methods that are computationally less demanding than LMG and PMVD and decompose R^2 into non-negative summands. As the section title suggests, these six proposals reduce to three different ones: it will be shown here that the proposals by Gibson[30] and Zuber and Strimmer[31] coincide; the coincidence between the proposals by Fabbris[1], Genizi[2] and Johnson[3] was e.g. pointed out by Nimon and Oswald[41]. All three methods are based on two different representations of the same optimum orthogonalization of the **X** matrix: an orthogonalized matrix **Z** may be either obtained from singular value decomposition as introduced by Johnson[60] or from standardization by the inverse of the unique symmetric square root of the correlation matrix. Assuming full column rank of the normalized matrix **X**, the orthogonalization from singular value decomposition starts from $\mathbf{X} = \mathbf{UDV}^T$ with an nxp matrix **U** with orthonormal columns, an orthogonal pxp matrix **V**, and a diagonal pxp matrix **D**. Johnson[60] proved that $\mathbf{Z}=\mathbf{UV}^T$ gives the set of p orthonormal vectors which is closest to the columns of **X**. His orthogonalization was used by Green et al.[46], Fabbris[1] and Johnson[3]. Gibson[30], Genizi[2] and Zuber and Strimmer[31] used the same orthogonalization in a different representation: they derived the optimum **Z** as $\mathbf{XR}_{XX}^{-1/2}$, where the power ½ denotes the unique symmetric square root of a matrix. The equality can be seen as follows: $\mathbf{R}_{XX} = \mathbf{X}^T\mathbf{X} = \mathbf{VDU}^T\mathbf{UDV}^T = \mathbf{VD}^2\mathbf{V}^T$, so that $\mathbf{R}_{XX}^{-1/2} = \mathbf{VD}^{-1}\mathbf{V}^T$. Thus, $\mathbf{Z} = \mathbf{X}\,\mathbf{VD}^{-1}\mathbf{V}^T = \mathbf{UDV}^T\mathbf{VD}^{-1}\mathbf{V}^T = \mathbf{UV}^T$.

The early Gibson[30] proposal – introduced with the hope to bring a resolution to the controversy between Hoffman[4,61] and Ward[8] – coincides with squared CAR scores by Zuber and Strimmer[31]: both

approaches simply use the squared coefficients c_j^2, j=1,…,p from regressing the normalized response vector **Y** on orthogonalized **Z** columns (see above) as surrogates for the corresponding normalized **X** values. These coincide with the explained variances and the R^2 values because of the standardization assumptions that have been made in this section. Instead of actually conducting these regressions, Zuber and Strimmer[31] showed that the c_j^2 can be obtained by squaring the components of

$$\mathbf{R}_{XX}^{-½}\mathbf{R}_{XY}. \tag{9}$$

In case of relevant correlations among the **X** variables, the **Z** variables can be far from being good representatives for the corresponding **X** variables – the principle has been visualized by Thomas et al.[62] for two dimensions (their Figure 1). Green et al.[46] therefore proposed to modify the c_j^2 values from Gibson[30] by relating the **Z** variables back to the **X** variables. They proposed to add two further steps:

- Calculate the squared coefficients from regressing the **Z** columns on the **X** columns, and obtain the proportions of the j-th original variable in the sum of squared coefficients for each **Z** column.
- Obtain the R^2 contribution of the j-th original variable as a weighted sum of squared coefficients c_j^2 from the Gibson approach with the proportions from the previous bullet as the weights.

Their proposal was criticized e.g. by Fabbris[1], who was among the first to note that it is not advisable to use the squared coefficients from regressing the **Z** columns on the **X** columns in case of correlated **X** columns. As the proposals by Fabbris[1], Genizi[2] and Johnson[3] coincide (see Nimon and Oswald[41]), their simplest representation by Johnson[3] is used for explaining the improvement on the Green et al. proposal: it simply regresses **X** on **Z** instead of **Z** on **X**. All other post processing steps for the c_j^2 remain the same. Alternatively to the weighted sum approach, the Fabbris / Genizi / Johnson decomposition can also be written as

$$((\mathbf{R}_{XX}^{½} \circ (\mathbf{1}_p^{\mathrm{T}} \otimes (\mathbf{R}_{XX}^{-½}\mathbf{R}_{XY})))\mathbf{1}_p)^{\mathrm{T}}, \tag{10}$$

where ° denotes the element wise product, ⊗ the Kronecker product, and $\mathbf{1}_p$ a column vector of p ones. The representation (10) was derived from the Zuber and Strimmer[31] paper.

Among the three different methods of this section, the present author considers the Fabbris / Genizi / Johnson method most convincing. Johnson[3] justified his proposal as an approximation to LMG and noticed that it often yields results quite similar to LMG. Thomas et al.[62] proved that it even algebraically coincides with LMG for the two regressor case. In spite of this, and conceding that the two are often close, they criticized the Fabbris / Genizi / Johnson metric as theoretically flawed and thus recommended that it should no longer be used. While the present author considers the use of LMG / dominance analysis as preferable, if feasible, Thomas et al.'s[62] fundamental condemnation of the Fabbris / Genizi / Johnson metric appears to be rooted in their assumption that Johnson[3] suffered under a naïve oversight when proposing his relative weights. The example in the next section will shed further light on the three approaches of this section in comparison to LMG / dominance analysis and PMVD.

Most methods discussed in this section are implemented in R package **relaimpo** (Grömping[54]; metrics "genizi" and "car" added in 2010): The Gibson[30] proposal – equivalent to squared CAR scores – is implemented under the name "car", using underlying work by Zuber and Strimmer[31] (package **care**), while the Fabbris / Genizi / Johnson approach is implemented under the name "genizi". To the author's knowledge, the Green et al.[46] approach is not available in R software. The R package **care** by

Zuber and Strimmer[31] provides regularization methods for situations with many variables. The R package **relaimpo** does not attempt to cover the many variables situation – nevertheless, all implemented metrics can in principle be applied to a regular covariance matrix that can e.g. be obtained by the regularization methods provided in R package **corpor** (Schäfer et al.[63]) in large variable situations (but LMG and PMVD will fail for resource reasons with many variables). Given that the Fabbris / Genizi / Johnson method has been advocated as an approximation to LMG, this might be the method of choice, if LMG is not feasible for resource reasons. Usage of that method is also recommended by Bi[43] in his review of relative importance methods from the sensory perspective.

EXAMPLE FOR THE VARIABLE IMPORTANCE METRICS IN THE LINEAR MODEL

In this section, all metrics discussed so far are exemplified, applying them to a historic socio-demographic data set of 182 provinces from Switzerland (Office of Population Research at Princeton University[64]) that was also used in Grömping[65]. The response variable, a Fertility index, is investigated dependent on five regressors: the percentage of males living on agricultural jobs, the percentage of draftees who obtained the best mark in an army exam, the percentage of draftees with more than the minimum level of education, the percentage of catholics in the province, and the percentage of live born infants who lived less than a year. A small portion of this dataset (47 provinces) ships with the R software. As the data set is used as an example only, it has neither been attempted to build a good model nor have small inconsistencies between the small built-in data set on 47 provinces and the larger data set been investigated.

Table 2 and Figure 1 show results from a linear model analysis of these data; all variables have been included linearly only, contrary to Grömping[65], where Agriculture and Catholic were allowed a quadratic term. This decision has been made in order to easily compare all metrics, including the simple ones for which grouped consideration of several regressors is not easily possible (e.g. squared standardized *b*'s). The R^2 of the model with only linear effects is more than ten percent points lower than that of the model that adds squared terms for Agriculture and Catholic, and – not surprisingly – leads to much lower (overall) shares for those two regressors. The analysis exemplifies the severity of the order dependence of metric (vii) (Sequential SS); as this is completely unreasonable, the metric is excluded from the further discussions. All other metrics agree on "Education" being the most important predictor, and most metrics place "Examination" second, however with vastly varying numeric allocations. Most metrics also place "Infant.Mortality" last. Numerically, among the variance decomposition metrics, PMVD and Pratt are quite close and allocate a large share to Education and a much smaller share to Examination, LMG and Fabbris / Genizi / Johnson are quite close and allocate a larger share to Examination, mainly by reducing the Education share, Gibson / CAR scores are in between these two and might be considered a compromise of the two perspectives, and the Green method is somewhat off (and has, in the author's opinion, been justly criticized by Johnson as being flawed).

TABLE 2 ABOUT HERE
FIGURE 1 ABOUT HERE

VARIABLE IMPORTANCE FOR PARAMETRIC NON-LINEAR MODELS

Theil and Chung[17] argued that the idea of averaging over orderings need not be limited to the linear model and R^2 or the variance, but can be extended to other scenarios. As an example, they proposed to apply the idea to information as a criterion, which they considered suitable for averaging because of its

additivity. Chevan and Sutherland[6] followed suit in extending the averaging over orderings to general goodness of fit measures in arbitrary multiple regression models. They introduced the term "hierarchical partitioning" for their proposal. Put simply, they proposed to choose an arbitrary goodness-of-fit metric G for the full model vs. the null model – for example R^2 or difference of deviances – and to allocate to the j-th regressor the unweighted average over orderings of the order-dependent contributions of the variable to G, analogously to formula (6). These allocations are then considered the "independent contributions" I_j of the j-th variable. The sum of all independent contributions is the overall goodness of fit of the full model. Additionally, the goodness of fit of the model with only the j-th variable vs. the null model is defined as the variable's overall contribution R_j, and the difference $J_j = R_j - I_j$ is called the "joint contribution" of the j-th variable. In addition to the above, Chevan and Sutherland[6] proposed to deep dive the joint contributions in order to understand the interplay between variables. Thus, hierarchical partitioning is similar to the afore-mentioned dominance analysis, but generally applicable to all models which offer any goodness of fit metric. LMG is a special case of hierarchical partitioning for linear models, with goodness of fit measure R^2. For generalized linear models, the R package **hier.part** (Walsh and MacNally[56]) implements the hierarchical partitioning approach, however without full detail on the joint contributions; note that the package can be relied upon to work correctly for up to 9 regressors only (and does not work at all for more than 12 regressors).

Retzer, Soofi, and Soyer[66] proposed to define importance in terms of the "information provided by a predictor for reducing the uncertainty about predicting the outcomes of the response variable". As information is a concept present in all statistical models, they suggested this approach as a unifying bracket for all types of regression situations and presented some example analyses in these terms. Apparently, however, this concept has not yet found its way into statistical practice. The author is not aware of any publicly available implementation in statistical software.

VARIABLE IMPORTANCE IN MACHINE LEARNING METHODS

In addition to the parametric regression methods, there are some variable importance methods for machine learning tools. In machine learning, the reason for the development of a "variable importance" feature was only indirectly the interest in variable importance itself; the more important reason was the motivation to take a look into the "black box" behind the methods – as the methods do not usually provide an explicit model equation, variable importance is a way to at least provide some idea about what drives the model. Here, we will briefly discuss variable importance for random regression forests (Breiman[67]; Strobl et al.[68]), as these can be understood as variance decomposition methods in a broader sense and have been investigated in connection with linear model variable importance by Grömping[65] in some detail. To the author's knowledge, forests and trees are the machine learning methods for which variable importance is best-researched. Nevertheless, some approaches have been considered for other methods, e.g. for neural networks (Gevrey et al.[25], as implemented in the R package **caret** by Kuhn[26]). Package **caret** implements a large number of variable importance methods – not all of them reasonable, e.g. absolute values of t-statistics for linear models – but many of them useful.

Random forests (and other machine learning tools) can be applied to all kinds of regression applications, among them categorical or survival response data. Different data types ask for different measures for entropy or variability; for example, the Gini impurity is often used for categorical responses. For quantitative responses, an MSE criterion is used to measure the forest's inaccuracy. Random regression forests are based on regression trees, in which the splitting process is guided by minimization of within node variance (CART trees, Breiman et al.[69]) or by significance tests

(conditional inference trees, Hothorn et al.[70], Hothorn, Hornik and Zeileis[71]). In a random regression forest, a large number of trees is built from a random selection of a small number m_{try} of variables and a random selection from the observations – the number m_{try} is a tuning parameter. The forest's prediction is the average prediction from individual trees. Each individual tree predicts a step function, an average over many trees can approximate almost any functional form and can automatically account for interactions between regressors. Of course, there is no closed form expression for the prediction equation. The most widely used variable importance metric for regression forests is permutation-based MSE reduction (see e.g. Strobl et al.[68]), called "permutation importance" in the following. This criterion was also investigated by Grömping[65] in comparison to the linear model variance decomposition methods LMG and PMVD. The relevant methods are implemented in the R packages **randomForest** (Liaw and Wiener[72]) and **party** (function `cforest`, Strobl et al.[68]), where the former implements the CART-based forests (RF-CART), the latter forests based on conditional inference trees (RF-CI).

Permutation importance is constructed from the random forest MSE as follows: In each tree, the out-of-bag observations, i.e. the observations not used for creating the tree, can be used for assessing the MSE from this particular tree. The overall performance of the forest can be obtained by an appropriate combination of the MSE estimates from individual trees. The contribution of a particular variable is determined by randomly permuting the observations for that variable and assessing the difference between the prediction performance with the actual and the permuted variable values.

Grömping[65] found similarities between the variable importances in random regression forests and the variance decomposition methods LMG and PMVD: The classical RF-CART (Breiman[67]) yields variable importances that are more similar to LMG allocations, while the RF-CI advocated by Strobl et al.[68] yields variable importances that are more similar to PMVD allocations. As was mentioned before, Budescu[14] and Johnson and Lebreton[15] emphasized the need for reasonable variable importance methods to integrate the conditional and the marginal approach. Like PMVD (and also the Pratt decomposition), the variable importances from RF-CI are much closer to the conditional perspective, while RF-CART yields importances that are closer to the more marginal perspective of LMG. The simulation study in Grömping[65] showed that estimation of the more conditional variable importances – be it PMVD, Pratt or the conditional random forest – yields much more variable results than estimation of the more marginal variable importances (LMG, CART-based random forests). Besides the intended purpose – for example, importance for prediction purposes may be of a more conditional nature, while importance for explanatory or causal purposes may require a higher impact of the marginal perspective – interpretation of variable importance results should also account for the variability of the result. Software for LMG and PMVD offers a variability assessment via the bootstrap; in forests, variable importances are usually not bootstrapped.

THE FUTURE OF THE FIELD

Variable importance is still actively researched. It is safe to predict that there will never be an agreed unique allocation of importances in case of correlated regressors. There may not even be a unique accepted definition of what variable importance is about. Nevertheless, interest in variable importance is large in many fields, as e.g. evidenced by the investigation conducted by Kruskal and Majors[35] and by many papers from various fields since.

Recently, Wang, Duverger and Bansal[73] attempted to combine dominance analysis with a Bayesian perspective in order to provide it with a theoretical basis. Likewise, the Pratt decomposition is revisited again and again by some authors who are fascinated by Pratt's axioms. This shows a need for

a fundamental answer to the variable importance question that the current rather heuristic best practices – like dominance analysis, LMG, PMVD, perhaps also the Fabbris / Genizi / Johnson method – have not been able to provide. In the author's view, the Pratt axioms are not a solution, but the fact that they are periodically revisited clearly indicates the understandable wish for a satisfactory theoretical foundation. Stufken[50] proposed that the game-theoretic background of LMG might provide a route to a better understanding of the properties of variable importance metrics. PMVD (Feldman[52]) also has a game-theoretic background as the proportional value. However, so far nobody has been able to substantially advance the understanding of variance decomposition methods based on game theory.

Variable importance methods have repeatedly been criticized for being atheoretical or unimportant, e.g. by Ehrenberg[74], who punned "The unimportance of relative importance". Certainly, crude relative importance methods like the ones discussed in this review will not suffice to deeply understand a subject matter. Exactly this is the reason why the author is wary about the more detailed aspects of dominance analysis or Chevan and Sutherland's[6] proposal. As Chevan and Sutherland justly stated – "Poorly specified models will not be improved by hierarchical partitioning or by any other partitioning method. Inadequate theory is the result of poor theorizing […] Statistical techniques do not build theory – theoreticians do." Nevertheless, the use of variable importance assessments, even from crude models, can provide ideas to the researcher, and, again citing Chevan and Sutherland[6] "need not of necessity lead to poorly specified models".

Usage in crude models has also been the reason for which Grömping[53] argued that the exclusion axiom brought forward by Feldman[29] is inadequate – at least for explanatory purposes – in case of correlated regressors: if the model does not make assumptions about a causal structure among the variables, even a coefficient 0 in the model with all other regressors is compatible with a relevant causal contribution of the regressor. This is a specific aspect of the controversy between the marginal and the conditional perspective on relative importance that was met both for the simple metrics (e.g. raw correlation vs. t-test), variance decomposition metrics (LMG vs. PMVD) and for forest based metrics (CART forests vs. conditional forests). As pointed out in Grömping[65], a related conflict also occurs in variable importance for variable selection: identification of a small number of variables sufficient for good prediction of the response variable is best served by a conditional perspective, while identification of important variables for explanatory purposes / interpretation is served better by the marginal perspective.

As long as no fundamental solution to the variable importance problem has been found, the author's recommendation is to use the existing best practices: For variance (or generally goodness of fit) decomposition based importance assessment, LMG enhanced with joint contributions or dominance analysis / hierarchical partitioning without too much detail, in case of many variables the Fabbris / Genizi / Johnson decomposition as a surrogate for these can be considered best practices. For more conditionally-inclined researchers, PMVD – or in case of non-negative shares the Pratt metric – can provide an alternative in the linear model. The example suggested that the Gibson / CAR scores might provide a compromise between the marginal and the conditional perspective, even though the theoretical concept of simply relying on an orthogonalization of correlated **X** variables does not appear particularly convincing. The machine-learning based variable importances are also viable alternatives. With the availability of modern computers and implementation of the advanced and computer-intensive methods into statistical software (at least, many are available in the open source R software[26,31,54,56,68,72]), calculation simplicity has become much less of an issue. Nevertheless, for large problems, computing power is still a concern. Furthermore – depending on the software they are using –, less statistically inclined users will still rely on simple metrics that are immediately available from the output of their software. It would therefore be desirable for commercial software packages to pick

up on the developments in variable importance and not restrict themselves to simplistic metrics like t-statistics or standardized coefficients. While it is understandable that software producers hesitate to implement approaches before they have reached some agreement within the scientific community, in the case of variable importance, heuristics and ambiguity are likely to persist for a long time, if not for ever. Therefore, implementation of some recommended methods in the near future would be more than welcome in order to enable more users to choose a method by suitability rather than availability. In spite of this appeal for implementation of variable importance metrics, ~~Generally,~~ it is also recommended to maintain awareness of the limitations regarding the insights that can be gained from variable importance considerations in correlated regressor situations.

REFERENCES

1. Fabbris, L. (1980). Measures of regressor importance in multiple regression: An additional suggestion. *Quality and Quantity* **4**, 787-792.
2. Genizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica* **3**, 407-420.
3. Johnson, J.W. (2000). A heuristic method for estimating the relative weight of regressors in multiple regression. *Multivariate behavioral research* **35**, 1-19.
4. Hoffman, P.J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin* **57**, 116-131.
5. Pratt, J.W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In: Pukkila, T. and Puntanen, S. (Eds.): *Proceedings of second Tampere conference in statistics*, University of Tampere, Finland, 245-260.
6. Chevan, A. and Sutherland, M. (1991). Hierarchical Partitioning. *The American Statistician* **45**, 90-96.
7. R Core Team (2014). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.
8. Ward, J.H. (1962). Comments on ”The paramorphic representation of clinical judgment”. *Psychological Bulletin* **59**, 74-76.
9. Darlington, R.B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin **69**, 161-182.
10. Bring, J. (1996). A geometric approach to compare variables in a regression model. *The American Statistician* **50**, 57-62.
11. Thomas, D.R., Hughes, E. and Zumbo, B.D. (1998). On variable importance in linear regression. *Social Indicators Research* **45**, 253-275.
12. Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression.* Sage, Thousand Oaks, CA.
13. Achen, C.H. (1982). *Interpreting and Using Regression*. Sage, Newbury Park, CA.
14. Budescu, D.V. (1993). Dominance Analysis: A new approach to the problem of relative importance in multiple regression. *Psychological Bulletin* **114**, 542-551.
15. Johnson, J.W. and Lebreton, J.M. (2004). History and Use of Relative Importance Indices in Organizational Research. *Organizational Research Methods* **7**, 238 - 257.
16. Holgersson, H.E.T., Norman, T., Tavassoli, S. (2014). In the quest for economic significance: Assessing variable importance through mean value decomposition. *Applied Economics Letters* **21**, 545-549.
17. Theil, H. and Chung, C.-F. (1988). Information-theoretic measures of fit for univariate and multivariate linear regressions. *The American Statistician* **42**, 249-252.
18. Silber, J. H., Rosenbaum, P. R. and Ross, R. N. (1995). Comparing the Contributions of Groups of Predictors: Which Outcomes Vary with Hospital Rather than Patient Characteristics? *J. Amer. Statist. Assoc.* **90**, 7-18.
19. Firth, D. (2011). relimp: Relative Contribution of Effects in a Regression Model. R package version 1.0-3. In: R Core Team (2014).

20. Ortmann, K.M. (2013). A Cooperative Value in a Multiplicative Model. *Central European Journal of Operations Research* **21**(3), 561-583.
21. Land, M. and Gefeller, O. (2000). A multiplicative variant of the Shapley value for factorizing the risk of disease. In Parrone, F., Garcia-Jurado, I. and Tijs, S.: *Theory and Decision Library* **23** (*Game Practice: Contributions from Applied Game Theory*), 143-158.
22. Eide, G. E. and Gefeller, O. (1995). Sequential and average attributable fractions as aids in the selection of preventive strategies. *J. Clin. Epidemiol.* **48**, 645–655.
23. van der Laan, M. (2006). Statistical inference for variable importance. *International Journal of Biostatistics* **2**(1), 1-33.
24. Ritter, S.J., Jewell, N.P., Hubbard, A.E. (2014). R Package multiPIM: A Causal Inference Approach to Variable Importance Analysis. *Journal of Statistical Software* **57**(8), 1-29.
25. Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling* **160**(3), 249-264.
26. Kuhn, M. (2014). caret: Classification and Regression Training. R package version 6.0-30. In: R Core Team (2014).
27. Pedhazur, E.J. (1982, 2nd ed.). *Multiple regression in behavioral research: explanation and prediction*. Holt, Rinehart and Winston, New York.
28. Ortmann, K.M. (2000). The proportional value of a positive cooperative game. *Mathematical Methods of Operations Research* **51**, 235-248.
29. Feldman, B. (1999). The proportional value of a cooperative game. *Manuscript for a contributed paper at the Econometric Society World Congress 2000*. Downloadable at http://fmwww.bc.edu/RePEc/es2000/1140.pdf.
30. Gibson, W.A. (1962). Orthogonal predictors: a possible resolution of the Hoffman-Ward controversy. *Psychological Reports* **11**, 32-34.
31. Zuber, V. and Strimmer, K. (2011). High-Dimensional Regression and Variable Selection Using CAR Scores. *Statistical Applications in Genetics and Molecular Biology* **10**(1), Article 34.
32. Lindeman, R.H., Merenda, P.F. and Gold, R.Z. (1980). *Introduction to Bivariate and Multivariate Analysis*, Scott, Foresman, Glenview IL. (p.119ff)
33. Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician* **41**: 6-10.
34. Kruskal, W. (1987b): Correction to "Relative importance by averaging over orderings". *The American Statistician* **41**: 341.
35. Kruskal, W. and Majors, R. (1989). Concepts of relative importance in recent scientific literature. *The American Statistician* **43**: 2-6.
36. MacNally, R. (2000) Regression and model building in conservation biology, biogeography and ecology: the distinction between and reconciliation of 'predictive' and 'explanatory' models. Biodiversity and Conservation 9: 655-671.
37. MacNally, R. (2002) Multiple regression and inference in conservation biology and ecology: further comments on identifying important regressors. Biodiversity and Conservation 11: 1397-1401.
38. MacNally, R. & Walsh, C. (2004). Hierarchical partitioning public-domain software. *Biodiversity and Conservation* **13**, 659-660.
39. Soofi, E.S., Retzer, J.J. and Yasai-Ardekani, M. (2000). A Framework for Measuring the Importance of Variables with Applications to Management Research and Decision Models. *Decision Sciences* **31**, 1-31.
40. Lebreton, J.M., Ployhart, R.E. and Ladd, R.T. (2004). A Monte Carlo Comparison of Relative Importance Methodologies. *Organizational Research Methods* **7**, 258 - 282.
41. Nimon, K.F. and Oswald, F.L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods* **16**, 650 - 674.
42. Bi, J. and Chung, J. (2011). Identification of drivers of overall liking – determination of relative importances of regressor variables. *Journal of Sensory Studies* **26**, 245-254.

43. Bi, J. (2012). A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. *Journal of Sensory Studies* **27**, 87–101.
44. Lipovetsky, S. and Conklin, M. (2001). Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry* **17**, 319-330.
45. Grömping, U. and Landau, S. (2010). Do not adjust coefficients in Shapley value regression. *Applied Stochastic Models in Business and Industry* **26**, 194-202. DOI: 10.1002/asmb.773.
46. Green, P.E., Carroll, J.D. and DeSarbo, W.S. (1978). A new measure of regressor importance in multiple regression. *Journal of Marketing Research* **15**, 356-360.
47. Azen, R. and Budescu, D.V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods* **8**, 129-148.
48. Budescu, D.V. and Azen, R. (2004). Beyond Global Measures of Relative Importance: Some Insights from Dominance Analysis. *Organizational Research Methods* **7**, 341 - 350.
49. Christensen, R. (1992). Comment on Chevan and Sutherland. *The American Statistician* **46**, 74.
50. Stufken, J. (1992). On hierarchical partitioning. *The American Statistician* **46**, 70-71.
51. Shapley, L. (1953). A value for n-person games. Reprinted in: Roth, A. (1988, ed.): *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, Cambridge. (*game-theoretic background for lmg*)
52. Feldman, B. (2005). Relative Importance and Value. Manuscript (latest version), downloadable at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2255827.
53. Grömping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician* **61**, 139-147.
54. Grömping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software* **17**, Issue 1.
55. Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D.W., Slotow, R. and Hamer, M. (2014). Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. *Methods in Ecology and Evolution* **5**, 320-328.
56. Walsh C. & Mac Nally, R. (2013). The hier.part Package: Hierarchical Partitioning. (Part of: *Documentation for R: A language and environment for statistical computing*.) R Foundation for Statistical Computing, Vienna, Austria. URL: http://cran.r-project.org/web/packages/hier.part/hier.part.pdf.
57. Azen, R. (2003). Dominance Analysis SAS Macros. URL: www.uwm.edu/~azen/damacro.html.
58. Fickel, N. (2001). Sequenzialregression: Eine neodeskriptive Lösung des Multikollinearitätsproblems mittels stufenweise bereinigter und synchronisierter Variablen. Habilitationsschrift, University of Erlangen-Nuremberg. VWF, Berlin.
59. Fickel, N. (2003). Measuring Supplementary Influence by Using Sequential Linear Regression. *Mathematics Preprint Archive, Volume 2003, Issue 4, April 2003*, 554-573.
60. Johnson, R.M. (1966). The minimal transformation to orthonormality. *Psychometrika* **31**, 61-66.
61. Hoffman, P.J. (1962). Assessment of the independent contributions of predictors. *Psychological Bulletin* **59**, 77-80.
62. Thomas, D.R., Zumbo, B.D., Kwan, E. and Schweitzer, L. (2014). On Johnson's (2000) Relative Weights Method for Assessing Variable Importance: A Reanalysis. *Multivariate Behavioral Research* **49**, 329-338.
63. Schäfer, J., Opgen-Rhein, R., Zuber, V., Ahdesmäki, M., Silva, A.P.D. and Strimmer, K. (2013). corpcor: Efficient Estimation of Covariance and (Partial) Correlation. R package version 1.6.6. In R Core Team (2014).
64. Office of Population Research at Princeton University (no year given). Switzerland Socio-economic variables 1870 to 1930. Website of the Princeton European Fertiliy Project. Data downloadable at
http://opr.princeton.edu/archive/fileList.aspx?studyid=10&substudy=x
(file efswitz.dat1888).

65. Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression Versus Random Forest. *The American Statistician* **63**, 308-319.
66. Retzer, J.J., Soofi, E.S. and Soyer, R. (2009). Information importance of predictors: Concept, measures, Bayesian inference, and applications. *Computational Statistics and Data Analysis* **53**, 2363–2377.
67. Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.
68. Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics* **9**, 307.
69. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Chapman and Hall, Boca Raton.
70. Hothorn, T., Hornik, K., van de Wiel, M.A. and Zeileis, A. (2006). A Lego System for Conditional Inference. *The American Statistician*, **60**(3), 257–263.
71. Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.
72. Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News* **2**(3), 18-22.
73. Wang, X., Duverger, P. and Bansal, H.S. (2013). Bayesian inference of predictors relative importance in linear regression model using dominance hierarchies. *International Journal of Pure and Applied Mathematics* **88**, 321-339.
74. Ehrenberg, A.S.C. (1990). The unimportance of relative importance. *The American Statistician* **44**, 260.

Tables

Table 1: Requirements and metrics

Requirement		(i): b_j	(ii) or (iv): t_j, p_j, $r^2_{j,other}$	(iii): r_j^2	(v): $b_j^2 s_j^2/s_y^2$	(vi): $b_{j,st} r_j$	(vii): *Sequential R^2 increase*	Gibson[30] / CAR scores[31]	Fabbris[1] / Genizi[2] / Johnson[3]	LMG[32-34] / dominance analysis[14]	PMVD[29]
Section[1)]		A	A	A	A	A	A	C	C	B	B
Anonymity	(a)	X	X	X	X	X		X	X	X	X
First two moments only	(b)	X	X	X	X	X	[2)]	X	X	X	X
Invariant to linear transformations on individual variables	(c)		X	X	X	X	X	X	X	X	X
Pure noise does not change anything	(d)	X	X	X	X	X	X	X	X	X	X
Balance conditional and marginal	(e)					X[3)]		X[3)]	X[3)]	X[3)]	X[3)]
Proper decomposition	(f)					X	X	X	X	X	X
Proper decomposition for each orthogonal subgroup	(g)					X	X	X	X	X	X
Non-negative shares	(h)						X	X	X	X	X
Exclusion	(i)	X	X		X	X					X
Inclusion	(j)	X	X		X			X	X	X	X
Pratt m/n	(k)					X					
Pratt non-singular multivariate linear transform	(l)					X					

1) A: Section "Simple metrics for measuring relative importance in linear regression models"
   B: Section "LMG, dominance analysis and PMVD:
   computer-intensive methods related to game theory"
   C: Section "Gibson decomposition / CAR scores, Green et al. decomposition and Fabbris / Genizi / Johnson decomposition"
2) Criterion b) cannot be satisfied if anonymity is violated.
3) These criteria do incorporate conditional and marginal elements. However, they may still lean towards one or the other extreme.

Table 2: Variable importance metrics for a fertility index in 1888 Swiss socio-demographic data

Relative importance in % normalized to sum 100% for all metrics* (R^2: 50.2 %%)	*Agriculture*	*Examination*	*Education*	*Catholic*	*Infant mortality*
b's (not normalized)	*0.1254*	*–0.2667*	*–0.4859*	*0.0335*	*0.2109*
Joint contribution (not normalized)	*2.7*	*21.4*	*18.6*	*5.8*	*–0.1*
Squared semipartial correlations	16.2	9.6	61.9	9.8	2.5
Squared raw correlations	5.8	41.4	42.3	10.2	0.3
Squared standardized b's	7.1	13.5	72.7	5.6	1.1
Sequential SS, from left to right	11.5	72.5	12.4	3.1	0.6
Sequential SS, from right to left	3.6	2.6	73.8	19.6	0.5
Pratt	6.9	25.3	59.2	8.1	0.6
CAR scores / Gibson	6.3	34.2	51.1	7.8	0.6
Green et. al.	2.0	44.3	47.5	5.3	0.6
Fabbris / Genizi / Johnson	6.5	36.1	47.7	9.0	0.8
LMG	6.1	38.7	45.9	8.5	0.8
PMVD	4.8	23.5	64.0	7.1	0.6

*Normalization to 100% is not recommended for data analysis, but has been chosen here for better comparability of the metrics.
Calculations have been conducted with R package **relaimpo**; the Green et al. method has been implemented using modified code from Bi and Chung[42] for Johnson's method.
Data and code are available as supplementary material.

Figure 1: Visualization of the normalized metrics from Table 2

The metrics are ordered by increasing allocation to Examination.
Blue stands for metrics that fulfill the Exclusion criterion,
yellow/orange stands for the more marginally-inclined metrics.
Green stands for metrics in-between or un-typified,
grey for the sequential allocations.

Related Articles

Article ID	Article title
21	e.g., Experimental mathematics and computational statistics
124	Multivariate analysis
487	Random Forests

Kommentiert [JW1]: Please insert 1-3 related articles in WIREs Computational Statistics to which your article may usefully be linked. Double-click on the icon below to see a list of all titles.