# Sélection de biomarqueurs pour prédire la tendreté de la viande de boeuf basée sur l'importance des variables

Marie Chavent[1] & Marie-Pierre Ellies-Oury[2] & Jérôme Saracco[3]

[1] *CQFD - Inria Bordeaux Sud Ouest, IMB, Université de Bordeaux, France*
marie.chavent@math.u-bordeaux.fr
[2] *Bordeaux Science Agro, 1 cours du Général de Gaulle, 33175 Gradignan, France*
*INRA, UMR1213 Herbivores, 63122 Saint Genès Champanelle, France*
*Clermont Université, VetAgro Sup, UMR1213 Herbivores, BP 10448, 63000 Clermont-Ferrand, France*
marie-pierre.ellies@agro-bordeaux.fr
[3] *CQFD - Inria Bordeaux Sud Ouest & IMB,*
*ENSC - Bordeaux INP, 109 Avenue Roul, 33400 Talence, France*
jerome.saracco@ensc.fr

**Résumé.** Dans cette communication, nous proposons une méthodologie computationnelle fondée sur l'importance des variables afin de choisir le modèle de régression le plus adéquat (parmi différents modèles) et de sélectionner simultanément les variables explicatives les plus pertinentes. L'application concerne la recherche des protéines les plus fortement liées à un phénotype (tendreté de la viande de boeuf). Trois modèles de régression différents (paramétrique, semiparamétrique et non paramétrique) et des méthodes associées (régression linéaire, régression inverse par tranches (SIR), forêts aléatoires) ont été utilisés pour choisir le meilleur et pour sélectionner simultanément les protéines les plus pertinentes pour prédire la tendreté parmi un ensemble de $p = 21$ biomarqueurs potentiels. Dans le muscle semi-tendineux des jeunes taureaux, dix protéines pourraient tre considérées comme des biomarqueurs de tendresse, notamment les protéines de choc thermique (HSP70-1B et HSP20) mais aussi métaboliques ($\beta$-enolase 3 et lactate déshydrogénase b).

**Mots-clés.** Importance des variables, modèle de régression, biomarqueurs, régression linéaire, régression inverse par tranches, forêts aléatoires.

**Abstract.** In this communication, a bench test of regression models to select variables strongly linked to a phenotypes is provided. Three different regression models (parametric, semiparametric and nonparametric) and associated methods (linear regression, sliced inverse regression, random forest) were used in order to choose the best one and to select simultaneously the more interesting proteins to predict tenderness among a pool of 21 potential biomarkers. In the semitendinosus muscle of young bulls, ten proteins could

be considered as tenderness biomarkers, especially heat shock proteins (HSP70-1B and HSP20) but also metabolic ones ($\beta$-enolase 3 and lactate dehydrogenase b).

**Keywords.** Importance of variable, regression model, biomarkers, multiple linear regression, sliced inverse regression, random forests.

# 1 Statistical methodology

From a general point of view, regression analysis studies the relationship between a $p$-dimensional predictor $X = (X_1, \ldots, X_k, \ldots, X_p)$ and a numeric response variable $Y$. Several ways to modeling this link are available. They can be divided in three main families: parametric, nonparametric and semiparametric modeling.

The parametric approach assumes that the underlying link function relies on a finite number of parameters to be estimated. In practice, the problem is to consider the good parametric form of the link function between $Y$ and $X$. For instance, linear regression is a particular case of parametric regression:

$$Y = \beta_0 + \sum_{k=1}^{p} \beta_k X_k + \varepsilon,$$

where $\varepsilon$ is a random error. In nonparametric regression, the class of fitted function is enlarged in order to obtain greater flexibility:

$$Y = f(X) + \varepsilon.$$

This increased flexibility has however a price to pay. Nonparametric approach provides a less understandable model and suffers from the curse of dimensionality (the efficiency strongly decreases as $p$ of increases). To overcome this drawback, it is proposed to combine nonparametric estimation method with dimension reduction technique. For instance, semiparametric (single index) model assumes that the response variable only depends on a linear combination of the predictors:

$$Y = f\left(\sum_{k=1}^{p} \beta_k X_k\right) + \varepsilon.$$

In statistical modeling, the problem of model choice is usual and crucial. Moreover, whatever the type of regression model, it is also necessary to select the relevant predictors in X in the chosen model. In statistical literature, many methods for predictors selection exist and are often specific to the underlying model and estimation method. In this paper, a computational model-free and estimation-free way to tackle the problem of model choice, variable selection included.

Whatever the considered model and the available sample $S = \{(x_i, y_i),\ i = 1, \ldots, n\}$, it is possible to calculate predicted values $\hat{y}_i = \hat{f}(x_i)$ where $\hat{f}$ is the estimated link function.

A common computational way, based on variable importance ($VI$), to select relevant predictors in the model is described below. $VI$ measures rely on estimating the response variable with some perturbations of the predictors and computing the error due to these perturbations: for the $j$th predictor,

$$VI_j = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{(j)}\right)^2,$$

where $\hat{y}_i^{(j)}$ is the predicted value based on the sample where the values of the $j$th predictor are randomly permuted. If the $j$th predictor has an effect on $Y$, the random permutation of its values in the sample will affect the estimation of $Y$ and its measure $VI_j$ will take a high value. In order to have a suitable idea of the importance of the $j$th predictor, it is necessary to replicate $N$ times this procedure: boxplot and mean of the $VI_j$s measures are obtained. This procedure is naturally applied to all the predictors ($j = 1, \ldots, p$). Parallel boxplots of the $VI$s values can be plotted to compare visually the importance of each predictor and select the relevant predictors using the mean square error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

as baseline. Another way is to identify automatically the useful predictors: an approach is to detect a single change point position (see [6]; in mean and variance) in the ordered sequence of the $p$ means of the $VI$s values. In order to determine which model (including variable selection step based on $VI$ approach) is the most accurate, a train/test samples approach can be used. The given sample is randomly split in a train sample $S_{train}$ (around 80% of the observations) and a test sample $S_{test}$ (around 20% of the remaining observations). Each model is constructed using the train sample and is evaluated on the test sample via the associated MSE:

$$MSE_{test} = \frac{1}{n_{test}}\sum_{i\in S_{test}}(y_i - \hat{y}_i)^2,$$

where $n_{test}$ is the size of $S_{test}$. The smaller the $MSE_{test}$ is, the better is the model. This procedure is replicated $M$ times to soften the random choice of train/test samples and parallel boxplots of $MSE_{test}$ (one boxplot per model) are used to visually select the most relevant model in terms of $MSE_{test}$. Simultaneously information on the number of useful variables automatically selected must be taken into account; moreover, the occurrence of each variable $X_k$, $k = 1, \ldots, p$ in the final model is also informative to determine the most relevant ones.

**Numerical simulations.** The good numerical behavior of the proposed methodology for model selection (including a step based on importance of variables in order to select the

useful predictors) has been illustrated on simulation. Two regression models have been considered. A parametric (linear) regression model (denoted M1) and a semiparametric (single index) regression model (denoted M2) are used to generate simulated samples. Three estimations methods are compared in this simulation study:

- a parametric estimation method : multiple linear regression (denoted RegLin hereafter),
- a semiparametric estimation method: sliced inverse regression associated with kernel regression (denoted SIR hereafter),
- a nonparametric estimation method : random forests (denoted RF hereafter).

The objectives of this numerical study is twofold:

- choose the best modeling (parametric, semiparametric or nonparametric) using $MSE$ (mean square error) criterion based on train/test samples approach,
- identify/select in the model the useful five predictors based on variable importance ($VI$) approach.

Naturally, ReLin method is efficient for M1 and suffers for M2. SIR plus kernel smoothing are well adapted for M1 and M2 even if the linear link function of M1 is nonparametrically estimated by the kernel regression of $Y$ given the estimated index. RF approach is purely nonparametric, this lack of dimension reduction step appears to be problematic in large dimensional spaces (when such a dimension reduction space exists, which is the case for regression models M1 and M2).

# 2 Application to tenderness of beef and molecular biomarkers

Tenderness is the most important attribute of beef quality, and its wide inconsistency is a major problem for beef industry. Thus, for many years, some researches were focused on tenderness determinism with the aim of better explaining and better predicting this parameter, thanks to the quantification of the abundance of molecules such as proteins (see [1], [2], [3], [4]). Several candidate proteins [dealing with apoptosis, oxidative stress, cytoskeletal proteins, proteolysis, oxidative metabolism, glycolytic metabolism, heat shock proteins and transport and signaling (see [1], [2], [5]) were identified as significantly linked with tenderness. Biomarkers were developed since the earlier methods of tenderness evaluation namely sensory panels as well as shear force methods are destructive, time consuming and ill-suited in routine as they require removing a piece of steak from the carcass to perform the measurement, hence leading to carcass depreciation. Thus, the

final aim of these researches is to select in a list of molecular biomarkers the ones that could be used to predict and predict meat quality.

To achieve that goal, most of previous papers are based on parametric methods like multiple linear regression or partial least squares regression.

To illustrate on real dataset the proposed methodology (implemented in R), multiple linear regression (RegLin; see [7] for instance) (parametric modeling), sliced inverse regression (SIR; see [8] for instance) (semiparametric modeling) and random forests (RF; see [9] for instance) (nonparametric modeling) are used. Any other statistical models and estimation methods could be used. More details and numerical results on simulated samples for various regression models can be found in supplementary information.

The selection among the three models was illustrated on experimental data obtained on $n = 71$ young bulls coming from the EU FP6 Integrated Project ProSafeBeef (FOODCT-2006-36241). The aim was to select among $p = 21$ muscular biomarkers (characterized by their abundances using the Dot-blot technique; see [10]) those that could predict the toughness of cooked m. Semitendinosus (evaluated instrumentally by Warner-Bratzler shear force using an Instron 5944).

The three models give equivalent results for mean square error. Nevertheless, the linear regression appears more variable and seems to suffer more when the replicating of a large number of train/test is done. When considering the number of selected biomarkers, the SIR method appears absolutely non selective as the best model to predict meat toughness remains 20 biomarkers on 21 in more than 70% of cases. The number of selected biomarkers is quite variable for random forest method, going from 2 to 20 biomarkers. Among 6 and 15 biomarkers are selected with the linear regression method. These two methods give similar results when looking at the biomarkers that were most often selected. The linear regression method being more adapted to suggest a predictive equation than the random forest one, and these two methods giving furthermore similar results, it could be proposed to keep this method for the present database. The selected biomarkers are respectively $\alpha$B-crystallin, heat shock protein 70.1B [HSP70.1B], $\beta$-enolase 3 [ENO3], lactate dehydrogenase b [LDHb], superoxide dismutase [SOD1], heat shock protein 20 [HSP20], peroxiredoxin6 [PRDX6], $\alpha$-actinin2, heat shock protein 40 [HSP40] and myosin heavy chain-I [MyHC-I] (when considering only the variables that are present in most than 60% of the cases). We demonstrated the importance of heat shock proteins such as $\alpha$B-crystallin, HSP 70-1B and HSP20 that were selected in 100% of randomization. We also confirmed the role of ENO3 and LDHb, involved in glycolytic metabolism (see [11]). Taken together, the 10 most important variables lead to a model with a multiple R-squared of 0.38, the variables with a positive significant link with toughness being HSP70-1B, ENO3 and HSP20, whereas LDHb and $\alpha$B-crystallin were found significantly negatively linked (PRDX6, HSP40 and $\alpha$-actinin2 being also negatively linked to toughness but non significantly in the present regression model). In this study, $\alpha$B-crystallin but also HSP 40 could be considered as tenderness biomarkers, whereas HSP70 and HSP20 are confirmed to be toughness biomarkers. The HSP70 are one of the most important HSP for maintaining

structural, ultrastuctural and functional properties of skeletal muscle. They are chaperone proteins that could inhibit meat tenderization by blocking protein-protein interactions of an important number of target proteins (see [12]). Moreover, HSP20 and HSP70 are known to inhibit apoptosis by sequestering pro-apoptotic factors such as Bax (see [13], [14]). Apoptosis has been identified as a tenderization enhancer just after slaughtering (see [15]), notably by protein proteolysis done by caspase. PRDX6, involved in redox reduction of the cell notably during oxidative stress, was previously found to be associated with tenderness mainly of the Semitendinosus muscle, but in opposite directions, depending on the breed (see [16]).

# 3 Conclusion

The originality of this work remains in the computational approach used especially concerning the method of variable selection that is not based on statistical inferences. It appears that whatever the model tested in the present work, the most important proteins to predict meat tenderness are the same, and are classified in the same order. Thus, we can conclude that the combinations of model choice and variable selection are robust. Nevertheless, with a multiple R-squared of 0.38, it appears that the biological mechanisms, when all proteins are taken together, depend on highly regulated mechanisms remaining unknown.

# Bibliographie

[1] Gagaoua, M., et al. (2015). J. Proteomics 128, 365-374.
[2] Moloto, W., et al. (2017). International Journal of Agriculture Innovations and Research 6, 467-1472.
[3] Ouali, A., et al. (2013).Meat Sci. 95, 854-870.
[4] Picard, B., et al. (2014). J Agric Food Chem., 62, 9808-9818.
[5] Gagaoua, M., et al. (2017). Meat Sci. 134, 18-27.
[6] Killick, R., & Eckley, I. A., (2014). Journal of Statistical Software 58, 1-19.
[7] Krzywinski, M. & Altman, N. (2015). Nat. Methods 12, 1103-1104.
[8] Li, K. C. (1991). Journal of the American Statistical Association 86, 316-342.
[9] Altman, N. & Krzywinski, M. (2017). Nat. Methods 14, 933-934.
[10] Guillemin, N., et al. (2009). Journal of Physiology and Pharmacology 60, 91-97.
[11] Guillemin, N.P., et al. (2012). Int. J. Biol. 4, 26.
[12] Houry, W.A., et al. (1999). Nature 402, 147-154.
[13] Beere, H.M., Green, D.R. (2001). Trends Cell Biol. 11, 6-10.
[14] Concannon, C.G., et al. (2003). Apoptosis 8, 61-70.
[15] Ouali, A., et al. (20006). Meat Sci. 74, 44-58 (2006).
[16] Picard, B., et al. (2014). J Agric Food Chem. 62, 9808-9818.