Journal of Applied Statistics
Vol. 00, No. 00, January 2009, 1–15

RESEARCH ARTICLE

Linear regression with compositional explanatory variables

K. Hron[a*], P. Filzmoser[b] and K. Thompson[c]

[a] *Palacký University, Faculty of Science, 17. listopadu 12, CZ-77146, Czech Republic*
[b] *Vienna University of Technology, Institute of Statistics and Probability Theory, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria*
[c] *Vienna University of Technology, Institute for Discrete Mathematics and Geometry, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria*

(*Received 00 Month 200x; in final form 00 Month 200x*)

Compositional explanatory variables should not be directly used in a linear regression model because any inference statistic can become misleading. While various approaches for this problem were proposed, here an approach based on the isometric logratio (ilr) transformation is used. It turns out that the resulting model is easy to handle, and that parameter estimation can be done like in usual linear regression. Moreover, it is possible to use the ilr variables for inference statistics in order to obtain an appropriate interpretation of the model.

Keywords: mixtures; Aitchison geometry on the simplex; isometric logratio transformation; orthonormal coordinates

1. Introduction

Regression analysis belongs to the most important tools in statistical analysis. Its goal is to explain the response variable Y using known explanatory variables $x_1, \ldots, x_D$. A linear regression model can be written in terms of a conditional expected value as

$$\mathsf{E}(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D, \tag{1}$$

with unknown parameters $\beta_0, \ldots, \beta_D$ that need to be estimated, e.g., using the standard least squares method. This approach is fully reasonable when both the response Y and the covariates $\mathbf{x} = (x_1, \ldots, x_D)'$ carry absolute information (represented often by variables corresponding to physical units). In many practical situations, however, the explanatory variables describe rather relative contributions of the components on the whole. In such a case, the sum of the variables (parts) is not important and the only relevant information is contained in the ratios between the parts. Usually, such data (called in the following *compositional data* or compositions [1]) are represented in proportions or percentages and are characterized by a constant sum constraint (1 or 100, respectively). As a consequence, the sample

*Corresponding author. Email: hronk@seznam.cz

ISSN: 0233-1934 print/ISSN 1360-0532 online

DOI: 10.1080/0266476YYxxxxxxxx
http://www.informaworld.com

space of D-part compositions $\mathbf{x} = (x_1, \ldots, x_D)'$ is the simplex,

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)',\, x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}.$$

It can thus be seen that compositional data are by definition singular. In what follows, under singular data we understand data with a singular covariance matrix. The sum of the parts, κ, can in principle be chosen arbitrarily because only the ratios of the parts contain the relevant information.

Historically, there were several approaches for regression models with compositional explanatory variables. Considering compositional data as observations that sum up to one was a starting point for so called *experiments with mixtures* [20, 21]. Linear combinations of the parts as well as canonical polynomials are considered,

$$\mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i x_i, \quad \mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i x_i + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \beta_{ij} x_i x_j,$$

the latter mainly for avoiding the singularity of the compositions, arising from their constant sum constraint, with obvious extensions to higher-degree polynomials. The parameters are estimated by the usual least squares method, however, the models are usually characterized by bad conditionality, thus, often some biased alternative (like Ridge regression [14]) needs to be performed. With mixture models, there have been many attempts to provide interpretations for the parameters of the model. Obviously, the source of the difficulties with the interpretation lies in the constant sum constraint; it is impossible to alter one proportion without altering at least one of the other proportions.

A big advance in this area was achieved by Aitchison in the early eighties [1, 2] by introducing the *logcontrast*, i.e. a term of the form $\beta_1 \ln x_1 + \cdots + \beta_D \ln x_D$, where the coefficients fulfill the condition $\beta_1 + \cdots + \beta_D = 0$ in order to follow the definition of compositions. Specifically, if any two coefficients in the linear combination are 1 and -1 and the remaining coefficients are zero, the logcontrast simplifies to a logarithm of a ratio, the so called *logratio*. Linear and quadratic *logcontrast models* (with the "logcontrast condition" $\sum_{i=1}^{D} \beta_i = 0$) are defined as

$$\mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i \ln x_i, \quad \mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i \ln x_i + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \beta_{ij} \left(\ln \frac{x_i}{x_j} \right)^2. \tag{2}$$

The parameters within logcontrast models are again estimated using the least squares method. The estimation of the parameters still leads to numerical difficulties because of the large number of parameters (in the quadratic case), and the constraint on the parameters that needs to be considered. The problem of the interpretation of the parameters remains, especially in more complex models.

Compositional data do not follow the usual Euclidean geometry, but they are described by the so-called Aitchison geometry on the simplex [4, 18]. The properties of this geometry are described in detail in various papers (see, e.g., [3, 7, 9, 10, 15]). Let us just mention that the Aitchison geometry follows the rules of a $(D-1)$-dimensional Euclidean space, and it is thus possible to construct an orthonormal basis (or a generating system) and to express the compositions therein. Hence, it is possible to find an isometric transformation to the usual Euclidean geometry. Because most statistical methods rely on the usual Euclidean geometry, the

compositions just need to be moved first isometrically from the simplex with the Aitchison geometry to the standard real space with the Euclidean one, using an appropriate *logratio transformation* that results in a real vector of logcontrasts. In the context of linear models with compositional explanatory variables the property of isometry is not directly necessary. Rather, the orthonormality of the resulting coordinates plays an important role. However, the orthogonality is a consequence of a regular isometric transformation (ilr transformation).

¿From the point of view of the isometry, the way forward might be provided by the *centred logratio (clr) transformation* [1], defined for a composition $\mathbf{x}$ as

$$(y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)', \tag{3}$$

that results in a model

$$\mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i y_i.$$

The clr transformation corresponds to coordinates with respect to a generating system on the simplex. For this reason, the resulting clr variables are singular ($\sum_{i=1}^{D} y_i = 0$) and the regression parameters should thus be estimated using the theory of singular linear models [16]. In addition, also here the interpretation of the regression coefficients could be misleading. The reason is that each clr variable explains the logratios between the part in the nominator and all parts in the composition, including itself. Thus, the clr variables as a whole explain some ratios more than once (which is another reason for the resulting singularity); see the next section for further discussion. A further problem with the clr transformation is its so-called subcompositional incoherence: Taking a subset of the parts would result in a linear regression model which might be incompatible with the full model with clr variables. A subset would alter each clr variable because all the parts in the currently used subset are contained in the denominator of (3).

All problems with the above mentioned models are hidden in the fact that the standard (unconstrained) linear model is meaningful if and only if the compositional covariates are expressed in an orthonormal basis on the simplex (with respect to the Aitchison geometry). Such a basis is given by an *isometric logratio (ilr) transformation* [8], and this approach will be studied in detail in this paper. The next section provides the definition and some basic properties of the ilr transformation. Section 3 introduces the linear regression model using ilr coordinates, and Section 4 shows the advantages of the proposed regression model for testing hypotheses about the influence of the relative contributions for explaining the response variable. The approach is illustrated with a data example in Section 5, and a discussion concludes the paper.

2. Properties of the isometric logratio transformation

An isometric logratio transformation seems to be the only way to achieve a regression model without the need for constraints on the parameters, and with a meaningful interpretation of the unknown parameters. The idea is to construct an orthonormal basis on the simplex, and to use the new coordinates in a standard linear regression model. Naturally, there are several ways to construct such

a basis, and in the context of compositional data this can be done with a method called sequential binary partitioning [9]. The result are coordinates that can be interpreted in terms of the involved compositional parts; a pre-knowledge on the studied problem usually leads to their better understanding [5, 6]. One choice which is used in different contexts ([8, 11, 15]) results in a $(D-1)$-dimensional real vector $\mathbf{z} = (z_1, \ldots, z_{D-1})'$, where the components are defined as

$$z_i = \sqrt{\frac{D-i}{D-i+1}}\, \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}},\ i = 1, \ldots, D-1. \tag{4}$$

The inverse transformation of $\mathbf{z}$ to the original composition $\mathbf{x}$ is then given, before closure, by

$$\begin{aligned}
x_1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1\right), \\
x_i &= \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} z_i\right), i = 2, \ldots, D-1, \\
x_D &= \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j\right).
\end{aligned}$$

The variable z_1 in (4) represents all the relevant information about the compositional part x_1, because it explains all the ratios between x_1 to the other parts of $\mathbf{x}$ [9, 13, 15]. Further, it is easy to see that if we were to permute the parts $x_2, \ldots, x_D$ in (4), the interpretation of z_1 remains unaltered. Even more, note that the interpretation of z_1 holds also when the remaining balances are chosen arbitrary according to a sequential binary partition of the subcomposition $x_2, \ldots, x_D$ [9] or another choice of the orthonormal basis on the simplex; however, the presented form of the balances seems to be most straightforward. Note also that there exists the linear relation $y_1 = \sqrt{\frac{D}{D-1}} z_1$ between the first clr and ilr variables, and thus the same interpretation as for z_1 holds also for y_1. Obviously, we cannot conclude that z_2 explains all the relative information about x_2 (in contrast to the second clr variable), because the part x_1 is not contained therein. In fact, z_2 explains the remaining ratios concerning x_2 (analogously for $z_3, \ldots, z_{D-1}$), and consequently, each logratio is uniquely explained by one ilr variable.

It is now straightforward to construct another orthonormal basis where the first ilr coordinate explains the compositional part we are interested in: It is sufficient to permute the indices in formula (4) in such a way that the part of interest plays the role of x_1. Consequently, for focusing on the lth part, for $l = 1, \ldots, D$, we need to construct D different ilr transformations, where the D-tuple $(x_1, \ldots, x_D)$ in (4) is replaced, e.g., by $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D) =: (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$. This results in the ilr transformation

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}}\, \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}},\ i = 1, \ldots, D-1, \tag{5}$$

and obviously we have $z_i^{(1)} = z_i$ for $i = 1, \ldots, D-1$, see (4).

Note that two different ilr transformations, resulting as expressions of $\mathbf{x}$ in different orthonormal bases on $\mathcal{S}^D$, are orthogonal transformations of each other [8]. This fact is important in proofs concerning the invariance of the results of regression models on the choice of the orthonormal basis for the ilr transformation.

3. Linear regression with ilr coordinates

Taking the above considerations into account, a natural way for designing a linear model between Y and $\mathbf{x}$ is to use the ilr transformation $\mathbf{z}$ of the composition $\mathbf{x}$ by using Equation (4) for example. One would then obtain a standard multiple linear regression of Y on the explanatory variables $\mathbf{z} = (z_1, \ldots, z_{D-1})'$ as

$$\mathsf{E}(Y|\mathbf{z}) = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_{D-1} z_{D-1}. \tag{6}$$

The regression coefficients $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{D-1})'$ can be estimated by the least squares method, where no constraints need to be imposed on the regression coefficients. The intercept term γ_0 is directly related to the response Y, it is not connected to the choice of the orthonormal basis on the simplex. Since the remaining regression coefficients are directly connected to the ilr coordinates, their interpretation needs to be adjusted accordingly [9]. As mentioned above, see (5), we can consider the lth ilr basis, for $l = 1, \ldots, D$, leading to a regression model

$$\mathsf{E}(Y|\mathbf{z}) = \gamma_0 + \gamma_1^{(l)} z_1^{(l)} + \cdots + \gamma_{D-1}^{(l)} z_{D-1}^{(l)}. \tag{7}$$

Due to the orthogonality of different ilr bases, the intercept term γ_0 (as well as the model fit) remains unchanged. Since $z_1^{(l)}$ explains all the relative information about part $x_1^{(l)}$, the coefficient $\gamma_1^{(l)}$ can be assigned to this part. The remaining regression coefficients are not straightforward to interpret since the assigned regressor variables do not fully represent one particular part. Thus, the only way to interpret the role of each compositional part for explaining the response Y is to consider D different regression models according to (7) by taking $l \in \{1, \ldots, D\}$, and to interpret the coefficient $\gamma_1^{(l)}$, representing part $x_1^{(l)}$.

In some applications it is possible to use another sequential binary partition for constructing orthonormal coordinates, different from those used in model (7). These orthonormal coordinates are describing certain relations between the compositional parts, and they have a unique interpretation. Accordingly, when using them in the regression model (6), all coefficients have a direct interpretation. However, this approach usually assumes a deeper a-priory knowledge of the underlying data and the relations between the variables.

Finally, note that comparing with the Aitchison's log-contrast models (2), each ilr variable is a log-contrast as well.

Having a sample with n observations of the response and of the explanatory variables, $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$, we can write the sample version of the linear model (6) as

$$Y_i = \gamma_0 + \gamma_1 z_{i1} + \cdots + \gamma_{D-1} z_{i,D-1} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{8}$$

where the explanatory variables $\mathbf{z}_i = (1, z_{i1}, \ldots, z_{i,D-1})'$ represent the ilr transformation of $\mathbf{x}_i$ (including 1 for the intercept). The usual basic assumptions on the random variables ε_i (uncorrelated, with the same variance σ^2) are assumed. With

the notation $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, the $n \times D$ design matrix $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$, and the error term $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, model (8) can be written as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{9}$$

The regression coefficients $\boldsymbol{\gamma}$ can be estimated by the least squares (LS) method, resulting in

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{10}$$

Note that a regression model with the ilr variables $z_1^{(1)}, \ldots, z_1^{(D)}$ would not be appropriate because it results in singularity (remember that $z_1^{(1)}, \ldots, z_1^{(D)}$ are just constant multiples of the clr variables). Concretely, here the corresponding design matrix of the regression model would have not full rank in columns and, consequently, it is not possible to use formula (10) for parameter estimation. Although one can employ alternative tools like singular value decomposition in order to obtain the desired estimators, ignoring the above facts would lead to serious restrictions of the quality of the estimation [12, 16].

Example: An illustration of different approaches is shown in Figure 1 for a small data set that can still be visualized. We consider the average expenditures of persons from European Union countries (year 2008) on food and for restaurants, which form the two explanatory variables. Obviously, these predictor variables are of compositional nature, since devoting more money to one part usually means that less money is left for the other part. The response variable is the GDP (gross domestic product) in the year 2008 of the same countries. Since Luxembourg has very extreme values, it will be excluded in this illustration. The data are available at `http://epp.eurostat.ec.europa.eu/`. The upper left panel of Figure 1 shows the original data in a 3-D plot, together with an LS regression plane resulting from a regression of the response on the original values of the predictors. Below is a plot of the response variable GDP versus the predicted response using this model. Obviously, the prediction quality is very poor, and the structure in the plot suggests that the linear model might not be appropriate. The middle upper and lower panel show the LS regression of GDP on the ratio of food to restaurants, as well as the response versus the prediction. A huge outlier (RO - Romania) is visible that is responsible for levering the LS line. Finally, the right upper and lower panel show the results of LS regression of GDP on the ilr transformed response variables. Note that according to (4), the ilr transformation for two parts x_1 and x_2 simplifies to $z_1 = \ln(x_1/x_2)/\sqrt{2}$. The linear regression model seems to be appropriate. Surely, this simple model is not very useful for predicting the GDP of the countries, but the example shows already the problems with the geometry of the space spanned by the explanatory variables. For the special case of two compositional predictor variables, the approaches with polynomials or logcontrast models would lead to comparable results in terms of model fit.

4. Inference in models with compositional explanatory variables

As in the standard multiple linear regression model, one is interested in testing hypotheses on the parameters $\gamma_0, \gamma_1, \ldots, \gamma_{D-1}$ in model (9). The required assumption for performing such tests is $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$, i.e., in addition to the previous assumptions, normal distribution is required. The significance of the individual

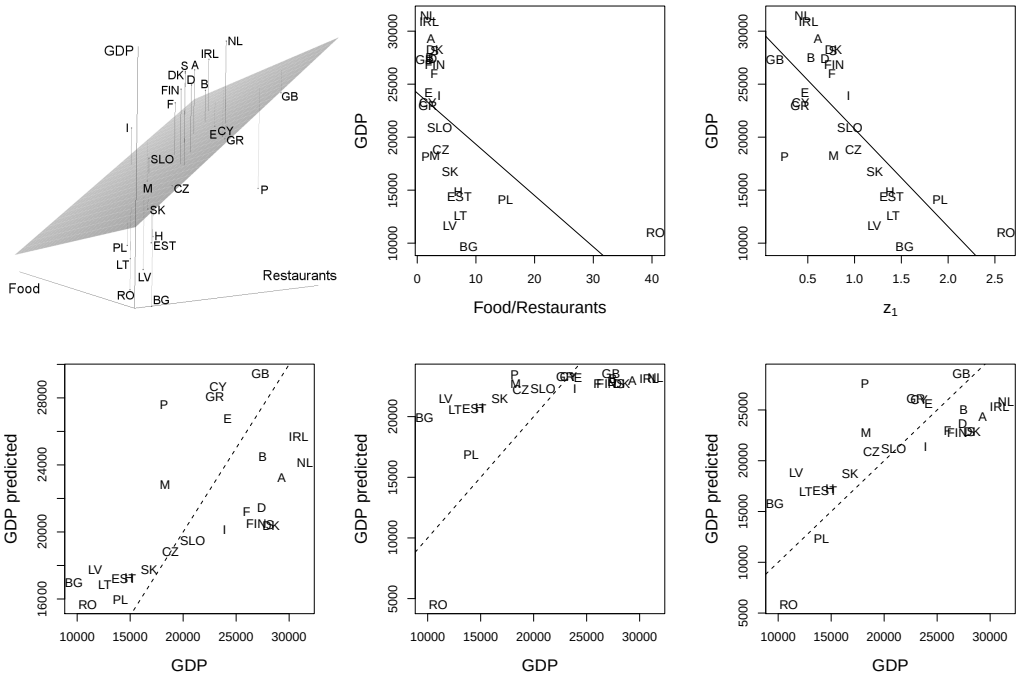

Figure 1. Illustration of different approaches for regressing the GDP on expenditures on food and restaurants measured in the countries of the EU. Shown are LS regression models for different transformations of the explanatory variables (upper plots), and the plot of the response versus the prediction (lower plots). Left column: regression for the original variables; middle column: regression on the ratio of food to restaurants; right column: regression on the ilr transformed explanatory variables.

regression parameters can be tested using the following statistics, see, e.g., [22],

$$T_0 = \frac{\widehat{\gamma}_0}{\sqrt{S^2\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{0,0}}}; \quad T_i = \frac{\widehat{\gamma}_i}{\sqrt{S^2\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{i,i}}}, \; i = 1, \ldots, D-1, \tag{11}$$

where $S^2 = (\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\gamma}})'(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\gamma}})/(n-D)$ is an unbiased estimator of the residual variance σ^2, and $\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{i,i}$ denotes the $(i+1)$-th diagonal element of the matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$, for $i = 0, \ldots, D-1$. Assuming the validity of the null hypothesis, T_0 and T_i follow a Student t-distribution with $n-D$ degrees of freedom.

As already discussed in Section 2, we are usually not interested in testing hypotheses for the parameters $\gamma_2, \ldots, \gamma_{D-1}$, because the corresponding predictor ilr variables are not straightforward to interpret. Thus, the focus is on the test statistics T_0 and T_1 which are used for testing the significance of the parameters γ_0 and γ_1, or, more generally, of the parameters γ_0 and $\gamma_1^{(l)}$, for $l = 1, \ldots, D$, see (7). Here special interest is on the parameter $\gamma_1^{(l)}$ belonging to the coordinate $z_1^{(l)}$ that carry all the relative information on the original part x_l. Consequently, the goal of the testing procedure is to find out, if a subcomposition (arising when x_l is omitted) of the given compositional covariate can replace the original composition in the regression model. Of course, one should be aware of the fact that, according to the definition of compositions, all the relevant information in a composition is contained in the ratios between the parts; as a consequence, the testing procedure (on significance of the chosen compositional part through the corresponding parameter $\gamma_1^{(l)}$) would produce different results when moving from the original compositional covariate to a subcomposition. In addition, the above considerations explain the key point, why the proposed choice of the balances (5) is of special i! mportance here: they are directly related to the inclusion or exclusion of a part in the explanatory

sub-composition. Finally, an alternative interpretation of the testing techniques is also to see which compositional parts have a significant influence on the response variable; the results of testing on γ_0 and $\gamma_1^{(l)}$, $l = 1, \ldots, D$, from models (7) using statistics (11) can be summarized in a table.

Here the main point that needs to be considered is the invariance of the test statistics T_0, T_1, used for testing the significance of the parameters γ_0 and $\gamma_1^{(l)}$ in (7), under relevant changes in the ilr basis. Although, for the sake of simplicity, in the following theorems only the established ilr transformations (5) are taken, the situation might be considered as a more general one. If we construct an orthogonal subspace to the first element of the basis, corresponding to coordinate $z_1^{(l)}$, it is possible to take also any other $D-2$ variables, coefficients of an orthonormal basis of the subcomposition $x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D$. Such a choice still guarantees that the resulting variables represent an isometric logratio transformation which enables to decompose the total variance of the composition [18]. However, note that it is possible to prove the following theorems also without the assumption of orthonormality of the basis of the subspace, in other words, for logcontrasts $z_2^*, \ldots, z_{D-1}^*$ that form an isomorphism between the simplex $\mathcal{S}^{D-1}$ and a $(D-2)$-dimensional real space. Here, however, it is no more possible to obtain a decomposition of the total variance of the composition as before. This general setting is followed in the Appendix, where the proofs are provided.

Theorem 4.1 Consider the linear model given in (7). T_0 is the test statistic for testing H_0: $\gamma_0 = 0$ against H_1: $\gamma_0 \neq 0$, and T_1 is the test statistic for testing $H_0^{(l)}$: $\gamma_1^{(l)} = 0$ against $H_1^{(l)}$: $\gamma_1^{(l)} \neq 0$, for a specific $l \in \{1, \ldots, D\}$.
(a) The test statistics T_0 and T_1 are invariant with respect to a change of the order of $x_2^{(l)}, \ldots, x_D^{(l)}$ in (5). As a consequence, the invariance on the choice of the coordinates holds also for the predicted values of the regression model;
(b) The test statistic T_0 is invariant with respect to a change of the order of $x_1^{(l)}, \ldots, x_D^{(l)}$ in (5).

Another important task for inference in regression analysis is whether the values of Y *at all* depend on values of the ilr coordinates $z_1, \ldots, z_{D-1}$. In other words, we want to test whether all the parameters γ_i, for $i = 1, \ldots, D-1$, are equal to 0. We may consider the statistic

$$F = \frac{1}{(D-1)S^2} \widehat{\boldsymbol{\gamma}}_*' \{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(-1,-1)} \widehat{\boldsymbol{\gamma}}_* \tag{12}$$

where $\widehat{\boldsymbol{\gamma}}_* = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_{D-1})'$ and $\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(-1,-1)}$ denotes that the first row and the first column were excluded from the matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$. If the null hypothesis holds, this test statistic follows the Fisher F distribution with $D-1$ and $n-D$ degrees of freedom.

Also here the invariance of F under the choice of the orthonormal basis is of primary interest. The proof of the following theorem can also be found in the Appendix.

Theorem 4.2 The test statistic F is invariant with respect to a change of the order of $x_1^{(l)}, \ldots, x_D^{(l)}$ in (5).

Finally, the quality of fit of the regression model (8) can be verified by the

coefficient of determination R^2, given as

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}, \tag{13}$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and $(\widehat{Y}_1, \ldots, \widehat{Y}_n)' = \mathbf{Z}\widehat{\boldsymbol{\gamma}}$ are predicted values of the response variable. As usual, the statistic R^2 takes on values in the interval $[0, 1]$, where values close to one indicate a strong linear relation between the explanatory variables to the response. Note that the above mentioned statistic F in (12) can be expressed in terms of the coefficient of determination R^2,

$$F = \frac{R^2}{1 - R^2}\frac{n - D}{D - 1}.$$

5. Examples

The following examples aim to provide more insight into the practical usage of regression with ilr-transformed regressor variables. The method has been implemented as function `lmilr` in the R package `robCompositions`, which can be freely downloaded from CRAN (the Comprehensive R Archive Network, `http://cran.R-project.org`), see [19]. As far as possible, a comparison with the standard regression approach based on the original data will be made.

5.1 *Relation between cancer and age structure*

We consider as the response variable Y the number of hospital discharges of inpatients on neoplasms (cancer) per 100 000 inhabitants (year 2007). This response is provided for the European Union countries (except Greece, Hungary and Malta) by Eurostat (`http://www.ec.europa.eu/eurostat`). As explanatory variables we use the age structure of the population in the same countries (year 2008). The age structure consists of three parts, age <15, age 15-60, and age >60 years, and they are expressed as percentages on the overall population in the countries. The data are provided by the United Nations Statistics Division (`http://unstats.un.org/unsd`). Regressing the response variable on the age structure alone may not be of primary interest for practitioners, and the resulting model might not be well suited for prediction purposes. Still, the model can give insights, because cancer is known to be dependent on the age of the persons. Moreover, using these regressor variables is technically interesting, because they sum up to 100% and thus result in perfect data singularity.

Figure 2 shows the original age variables drawn against the response variable. A regression in the original space would not make sense, and here it would not even be possible because of the singularity problem.

Table 1 shows the result from regression with ilr coordinates, in the form of the output of our R routine `lmilr`. The output is in the same style as the usual output from the R function `lm` for standard least-squares regression. The first block in the result listing refers to the parameters of the regression model, i.e. to the parameters γ_0, $\gamma_1^{(1)}$, $\gamma_1^{(2)}$ and $\gamma_1^{(3)}$ according to model (7) for $l = 1, 2, 3$. The columns correspond to the estimated regression parameters (`Estimate`), their standard errors (`Std. Error`), the value of the test statistic (`t value`) according to equation (11), and the corresponding p-value (`Pr(>|t|)`) for the test. It is important to emphasize that in Table 1 the results of three regression models are given, because in each of

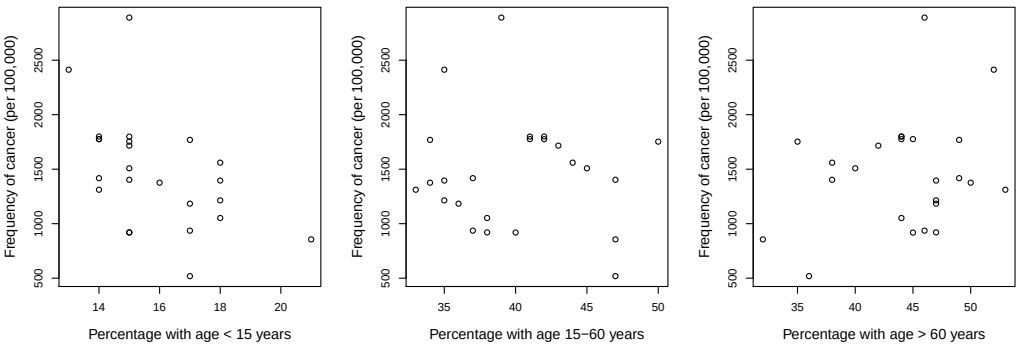

Figure 2. Plots of the hospital discharges of in-patients on neoplasms (cancer) per 100 000 inhabitants (vertical axes) versus the compositional parts defining the age structure in the 24 European Union member states.

them we focus just on the estimation (and testing) of parameters corresponding to the first coordinate $z_1^{(l)}$ and the intercept parameter. This is the main difference from the standard case, where each explanatory variable in the regression has a straightforward interpretation. Due to Theorem 4.1, the estimated value of the parameter γ_0 (and the corresponding test statistic T_0) can be computed using any of the above models, the same holds also for the test statistic F (Theorem 4.2).

The interpretation of the values of the estimated parameters $\gamma_1^{(l)}$ appears from the form of the coordinates $z_1^{(l)} = \sqrt{\frac{D-1}{D}}\ \ln \frac{x_l}{\sqrt[D-1]{\prod_{j\neq l} x_j}}$, whose values reflect besides the previously mentioned interpretation also the logarithm of the ratio between the part of interest and the geometric mean of the remaining parts in the composition (up to a constant that tends to 1 for an increasing number of parts D), i.e. the logratio of the part x_l and an average of the remaining parts. Consequently, the interpretation can be done in an analogous way as it is done in standard regression, namely, the value of the estimate indicates how much the response variable changes in average by a unit change of the above logratio representing the compositional part of interest.

Accordingly, the age groups < 15 and > 60 have significant contribution for explaining the response, where age group < 15 has negative influence and age group > 60 positive influence (see the signs of the estimated parameters). This outcome corresponds also to the intuition, because the higher the relative amount of elderly people, the higher the occurrence of cancer in the society will be, and for a higher relative amount of young people the cancer occurrence is expected to be lower. Further below in the listing of Table 1 there is information of the model

Table 1. Results from regression of the cancer variable on the ilr coordinates of the age structure. For detailed explanations see text.

```
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)       -357.2       693.5    -0.515     0.6119
ilr(Age <15)    -2275.8        872.8    -2.607     0.0164
ilr(Age 15-60)    965.7        733.7     1.316     0.2023
ilr(Age >60)     1310.2        581.7     2.252     0.0351

Residual standard error: 460.4 on 20 degrees of freedom
Multiple R-squared: 0.2694,     Adjusted R-squared: 0.1998
F-statistic: 3.872 on 2 and 21 DF,  p-value: 0.03704
```

fit. With about 27%, the coefficient of determination is rather low, which can also be expected because the age structure alone might just be indicative for the trend

of cancer, but will not be able to fully explain the response. Still the F statistic shows that explanatory variables and response have significant relation.

A naive approach for least-squares regression with the original explanatory variables would be to use only two of the three variables, arguing that due to the constant sum of 100%, any two variables contain the same information as all three age variables. Using the explanatory variables age <15 and age >60 years result in the output presented in Table 2. While the coefficient of determination is about the same as before, now only age <15 is significant. One could also use other two

Table 2. Results from regression of the cancer variable on the original variables age <15 and age >60 years. For explanations see text.

```
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)     3310.88     1494.57     2.215     0.0379
Age <15         -132.18       56.50    -2.340     0.0293
Age >60            5.69       19.87     0.286     0.7774

Residual standard error: 463.2 on 21 degrees of freedom
Multiple R-squared: 0.2605,     Adjusted R-squared: 0.1901
F-statistic: 3.699 on 2 and 21 DF,  p-value: 0.04206
```

explanatory variables, like age <15 and age 15-60 years. The result is shown in Table 3. It is interesting to note that the coefficient of determination is unchanged, but also that the coefficient for age >60 years in Table 2 is the same as that for age 15-60 years in Table 3, with a different sign, and that the corresponding test results in the same p-value. This relation becomes clearer by considering the model corresponding to Table 2, $\mathsf{E}(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$. Since the explanatory variables sum up to 100 (percent), we have $x_3 = 100 - x_1 - x_2$, and the model can be rewritten as $\mathsf{E}(Y|\mathbf{x}) = (\beta_0 + 100\beta_3) + (\beta_1 - \beta_3)x_1 - \beta_3 x_2$. Accordingly, the parameter estimates and the inference statistics become useless, and an interpretation of these results is highly incorrect.

Table 3. Results from regression of the cancer variable on the original variables age <15 and age 15-60 years. For explanations see text.

```
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)     3879.86     1101.46     3.522     0.0020
Age <15         -137.87       51.49    -2.678     0.0141
Age 15-60         -5.69       19.87    -0.286     0.7774

Residual standard error: 463.2 on 21 degrees of freedom
Multiple R-squared: 0.2605,     Adjusted R-squared: 0.1901
F-statistic: 3.699 on 2 and 21 DF,  p-value: 0.04206
```

5.2 *Relation between life expectancy and GDP groups*

We consider the average life expectancy at birth for women as the response variable Y, and different compositions of the gross domestic product (GDP) as explanatory variables. The data are taken from `http://unstats.un.org/unsd` for the European Union member states. Luxembourg is an outlier and will thus not be used here. Note that life expectancy at birth is an estimate of the number of years to be lived by a female newborn, based on current age-specific mortality rates. Obviously, there should be some influence of this characteristic on factors concerning the economic position of the member states. The GDP compositions are based on the international standard industrial classification (ISIC) of all economic activities,

and they are given for the following six categories: agriculture, hunting, forestry, fishing (ISIC A-B, `x1`); mining, manufacturing, utilities (ISIC C-E, `x2`); construction (ISIC F, `x3`); wholesale, retail trade, restaurants and hotels (ISIC G-H, `x4`); transport, storage and communication (ISIC I, `x5`); other activities (ISIC J-P, `x6`). The last category ISIC J-P contains activities on education, health and social work as well as other community, social and personal service activities. Accordingly, this category seems to be important for explaining variability of the response.

The original explanatory variables are expressed in percentages. Aside from the fact that we deal with compositions, this would cause a singularity problem when employing least-squares regression directly on the raw data. However, in order to make a comparison with a regression approach based on the raw data, we can multiply the values by the GDP per capita (expressed in USD) to obtain total amounts (per capita) devoted to the six GDP categories. Note that for the compositional approach such a multiplication does not have any influence on the final results. Table 4 shows the resulting output when applying least-squares regression to the ilr coordinates of the GDP categories, while Table 5 provides the results when using the raw untransformed data. For the ilr approach we obtain the expected result that `x6` is significant. In addition, for a significance level of 5%, also `x5` contributes significantly, which also seems to be plausible. On the other hand, according to Table 5 none of the original variables is significant. The coefficient of determination is even higher when using the raw data.

Table 4. Results from regression of the life expectancy for women on the ilr coordinates of the GDP categories.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	73.1982	1.9914	36.757	< 2e-16
ilr(x1)	-0.7836	1.0642	-0.736	0.470
ilr(x2)	-0.6999	1.3312	-0.526	0.605
ilr(x3)	0.5001	1.5373	0.325	0.748
ilr(x4)	0.0834	2.1237	0.039	0.969
ilr(x5)	-4.3847	2.0353	-2.154	0.044
ilr(x6)	5.2847	1.8184	2.906	0.009

```
Residual standard error: 1.798 on 19 degrees of freedom
Multiple R-squared: 0.5958,     Adjusted R-squared: 0.4947
F-statistic: 5.896 on 5 and 20 DF,  p-value: 0.001664
```

Table 5. Results from regression of the life expectancy for women on the original variables of the GDP categories.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.726e+01	1.097e+00	70.413	<2e-16
x1	2.208e-05	1.844e-05	1.198	0.246
x2	-1.281e-06	1.964e-06	-0.652	0.522
x3	-2.259e-06	6.106e-06	-0.370	0.715
x4	6.150e-06	4.144e-06	1.484	0.154
x5	-6.935e-06	7.959e-06	-0.871	0.394
x6	1.906e-06	1.178e-06	1.618	0.122

```
Residual standard error: 1.738 on 19 degrees of freedom
Multiple R-squared: 0.6412,     Adjusted R-squared: 0.5278
F-statistic: 5.658 on 6 and 19 DF,  p-value: 0.001631
```

Figure 3 shows the plots of the response variable versus the predicted response, using the ilr regression model (left) and the model for the original data (right). The structure in both plots is similar. It is interesting to see two groups, corresponding to the former so-called Eastern and Western European countries. The quality of

fit of both models is comparable, and from that point of view one could not give preference to any of the models. The big difference, however, is the interpretation of the models using the inference statistics shown in Tables 4 and 5: the results based on the original data are misleading because the data are not represented in the usual Euclidean space.

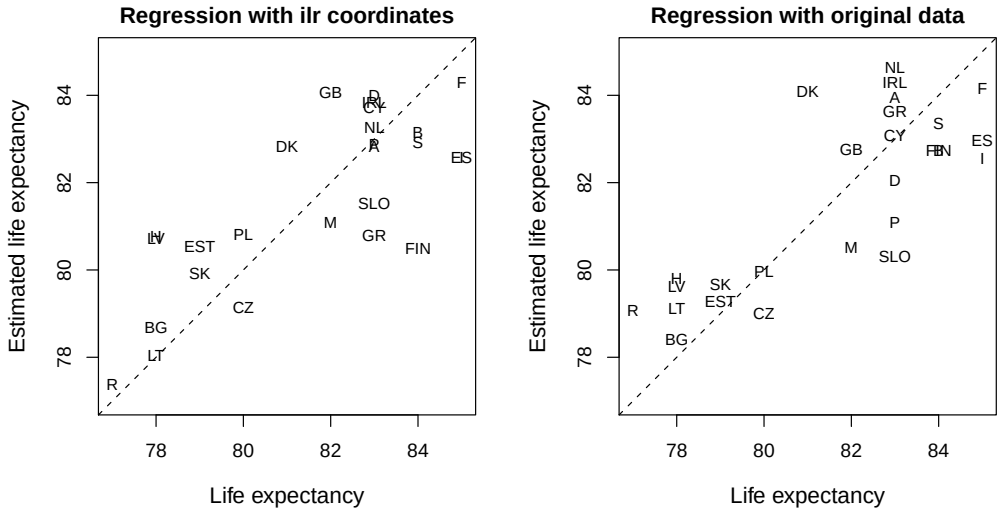

Figure 3. Response variable versus the predicted response, using the ilr regression model (left) and the model for the original data (right).

6. Discussion

Regression with compositional explanatory variables can be misleading if the original untransformed data are directly used in the regression model. The main problem with this approach is the geometry: Compositional data are not represented in the usual Euclidean space, but in the so-called Aitchison geometry on the simplex. Since there is a non-linear relation between the usual Euclidean and the Aitchison geometry, linear regression in the original data space would correspond to a non-linear regression in the Euclidean space, and vice versa. While the user might not be aware of this non-linearity by using the original explanatory variables, the resulting model might even have a better fit than a model for transformed data. The important point, however, is the interpretation of the model, and in particular the inference statistics for the regression parameters of the model, which is only valid if the data are represented in the appropriate Euclidean space.

For transforming the Aitchison geometry to the usual Euclidean space, the isometric logratio (ilr) transformation has the most preferable properties among all logratio transformations. In contrast to the centered logratio (clr) transformation, it avoids data singularity, which is an important issue in regression. A disadvantage of the ilr approach, however, is the difficulty with interpreting the newly constructed variables. Therefore, it is crucial how the ilr transformation is chosen. Here we propose to select the ilr basis in such a way that the first basis vector coordinate contains all the relative information about one particular compositional part. Hence, the parameter estimation and the inference statistic for this parameter fully refer to this part. The remaining ilr variables are used for regression, but they are not useful for the interpretation, because they cannot be assigned to one single compositional part. Thus, in order to enhance interpretability, another ilr basis

can be chosen, where again the first ilr coordinate contains all information about another specific part. This can be done for each explanatory variable. Since the different ilr transformations are orthogonal rotations of the corresponding bases, the fit of each model is exactly the same.

The proposed approach can be used in a much broader context: Since the contribution of each explanatory variable in the model can be estimated, this approach is suitable for variable selection techniques such as stepwise variable selection [22]. Furthermore, since the residuals are non-compositional values, not only least-squares estimation, but also other objective functions for the residuals could be used, like robust regression [17]. Outliers in the response variable can be treated as in the usual case, but outliers in the explanatory variables need to be treated from a compositional point of view. We will leave these topics for future research.

Acknowledgements The authors are grateful to helpful comments and suggestions of the referee. This work was supported by the Council of the Czech Government MSM 6198959214.

References

[1] Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London.
[2] Aitchison J, Bacon-Shone J (1984) Log contrast models for experiments with mixtures. Biometrika 71(2):323-330.
[3] Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio analysis and compositional distance. Math Geol 32:271-275.
[4] Billheimer D, Guttorp P, Fagan W (2001) Statistical interpretation of species composition. Journal of the American Statistical Association 96:1205-1214.
[5] Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (2006) Frequency distributions and natural laws in geochemistry. In Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V, eds (2006) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264:175–189.
[6] Buccianti A, Egozcue JJ, Pawlowsky-Glahn V (2008) Another look at the chemical relationships in the dissolved phase of complex river systems. Math Geosci 40: 475–488.
[7] Egozcue JJ (2009) Reply to "On the Harker Variation Diagrams" by J.A. Cortés. Math Geosci 41:829-834.
[8] Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Math Geol 35:279–300.
[9] Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Math Geol 37:795–828.
[10] Egozcue JJ, Pawlowsky-Glahn V (2006) Simplicial geometry for compositional data. In Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V, eds (2006) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264:145–160.
[11] Filzmoser P, Hron K, Reimann C (2009) Univariate analysis of environmental (compositional) data: Problems and possibilities. Science of the Total Environment, 407:6100–6108.
[12] Fišerová, E., Kubáček, K., Kunderová, E.: Linear statistical models - regularity and singularities. Academia, Praha, 2007.
[13] Fišerová E, Hron K (2010) On interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences, 43:455–468.
[14] Hoerl AE, Kennard R (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12:55–67.
[15] Hron K, Templ M, Filzmoser P (2010) Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis, 54:3095-3107.
[16] Kubáček L, Kubáčková L, Volaufová J (1995) Statistical models with linear structures. Veda, Bratislava.
[17] Maronna R, Martin RD, Yohai VJ (2006) Robust statistics: Theory and methods. John Wiley, New York.
[18] Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment, 15:384-398.
[19] R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
[20] Scheffé, H. (1958) Experiments with mixtures. Journal of the Royal Statistical Society - B, 20:344-360.
[21] Scheffé, H. (1963) The simplex-centroid design for experiments with mixtures. Journal of the Royal Statistical Society - B, 25:235-263.
[22] Varmuza K, Filzmoser P (2009) Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press, Boca Raton, FL, 2009.

7. Appendix

Proof of Theorem 4.1: Without loss of generality we set $l = 1$, and thus we refer to the notation used in Equation (8) with the ilr basis defined in (4).
(a) The change of the order of $x_2, \ldots, x_D$ in (4) corresponds to a change of the orthonormal basis on the simplex [8], i.e., the design matrix $\mathbf{Z}$ is multiplied from the right-hand side by a $D \times D$ orthogonal matrix

$$\mathbf{P} = \begin{pmatrix} 1 & & \\ & 1 & \\ & & \mathbf{P}_1 \end{pmatrix},$$

with values of one in the first two entries of the diagonal, a $(D-2) \times (D-2)$ orthogonal matrix $\mathbf{P}_1$, and zeros elsewhere. We obtain $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}_D$ (where $\mathbf{I}_D$ stands for identity matrix of order D). Even more, it is also possible to choose coordinates of a non-orthonormal basis to express the subcomposition $x_2, \ldots, x_D$ therein, like the additive logratio coordinates [1]. As a consequence, the matrix $\mathbf{P}$ looses the property of orthogonality, which, however, does not restrict the considerations below.

Using the relation

$$[(\mathbf{ZP})'\mathbf{ZP}]^{-1}(\mathbf{ZP})'\mathbf{Y} = \mathbf{P}^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{P}')^{-1}\mathbf{P}'\mathbf{Z}'\mathbf{Y} = \mathbf{P}^{-1}\widehat{\gamma}$$

we can see that the values of the estimates $\widehat{\gamma}_0$, $\widehat{\gamma}_1$ and S^2 as well as the first and second diagonal elements of the matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$ in (11) remain unchanged under the mentioned regular affine transformation. It immediatelly follows that for the statistics T_0 and T_1 the requested invariance is fulfilled. Finally, an obvious relation $\mathbf{ZPP}^{-1}\widehat{\gamma} = \mathbf{Z}\widehat{\gamma}$ holds, i.e., the fitted linear regression model is equal irrespective to the chosen basis (orthogonal or not).

(b) The proof processing is the same as before, where the matrix $\mathbf{P}$ is replaced by

$$\mathbf{Q} = \begin{pmatrix} 1 & \\ & \mathbf{Q}_1 \end{pmatrix},$$

with a $(D-1) \times (D-1)$ regular matrix $\mathbf{Q}_1$ (thus, $\mathbf{P}$ is a special case of the matrix $\mathbf{Q}$). The invariance of $\widehat{\gamma}_0$, S^2, of the first diagonal element of $(\mathbf{ZZ})^{-1}$, and consequently also of the statistic T_0 is obvious. □

Proof of Theorem 4.2:
To show that the F statistic is invariant under the choice of the orthonormal basis coming from a permutation of $x_1, \ldots, x_D$ in (4), or, more generally, under a change of any basis on the simplex (orthonormal or not), is proved analogously as in Theorem 4.1. The design matrix $\mathbf{Z}$ is again multiplied by the regular matrix $\mathbf{Q}$, thus from (12) we obtain that

$$\widehat{\gamma}_*'\mathbf{Q}_1\mathbf{Q}_1^{-1}\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(-1,-1)}(\mathbf{Q}_1')^{-1}\mathbf{Q}_1'\widehat{\gamma}_* = \widehat{\gamma}_*'\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(-1,-1)}\widehat{\gamma}_*$$

and the proof is complete. □