



City Research Online

City St George's, University of London

Citation: Ben-Gad, M. (2026). Compositional and Noncompositional Covariates in Multiple Linear Regression. Computational Statistics, 41(6), 115. doi: 10.1007/s00180-026-01774-y

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/37781/>

Link to published version: <https://doi.org/10.1007/s00180-026-01774-y>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).



Compositional and noncompositional covariates in multiple linear regression

Michael Ben-Gad¹ 

Received: 1 July 2025 / Accepted: 19 June 2026
© The Author(s) 2026

Abstract

Empirical studies in the social sciences often include covariates that are compositional in nature—vectors of shares that sum to a constant—alongside noncompositional covariates. A common approach to handling compositional data in linear regression is to omit one component to avoid perfect multicollinearity. I demonstrate why these coefficients (along with standard errors and t-statistics) can be highly sensitive to the choice of omitted category. Transforming the compositional data using additive logarithmic ratios (ALR) yields permutation-invariant regressions and permits counterfactual changes in the implied composition that remain within the simplex. Furthermore, I demonstrate that applying a simple scale factor to the ALR coefficients generates the coefficients and standard errors associated with isometric logarithmic ratios (ILR) for the variables of interest. Finally, using log-ratios does not exacerbate inherent problems of multicollinearity associated with compositional data. Economic growth regressions incorporating compositional and noncompositional covariates are used to illustrate.

Keywords Compositional data · Linear regression · Log-contrasts · Logarithmic ratios · Economic growth regressions

I would like to express my deepest gratitude to Vera Pawlowsky-Glahn, who carefully read through an earlier draft and provided invaluable feedback. Thank you also to Juan José Egozcue, John Luke Gallup, and Michael Greenacre, seminar participants at City St George's, University of London, and in particular to Mireia Jofre-Bonet, Giulia Faggio, Joseph Pearlman and Agne Suziedelyte.

✉ Michael Ben-Gad
Michael.Ben-Gad.1@citystgeorges.ac.uk

¹ Department of Economics, City St George's, University of London, Northampton Square, London EC1V 0HB, UK

1 Introduction

Multiple linear regressions often include data that can be expressed as coordinates in Euclidean space, alongside compositional data—i.e., vectors that describe parts of a whole, where the relevant information is relative, conveyed by the ratios between their strictly positive components, and representable as coordinates in a simplex. As the linear regression models presented in Sect. 2 typically include an intercept term, it is common in applied work to omit one component within a composition to ensure the matrix of explanatory variables has full column rank.¹ Sect. 3 provides formal proofs for the permutation invariance of fitted values, residuals, and coefficients on noncompositional covariates when one part is omitted and shows how the estimated coefficients and corresponding *t*-statistics for subcompositional components depend on which component is omitted.

The paper then turns to log-ratio representations. Section 4 shows that additive log-ratios (ALRs), developed by Aitchison (1982, 1986), provide a convenient single-regression representation of this log-contrast and remove the arbitrariness associated with the omitted-category formulation.

Section 5 considers isometric log-ratios (ILRs), developed by Egozcue et al. (2003). In Hron et al. (2012), recovering the coefficients of greatest substantive interest—those corresponding to the log deviation of a variable from the geometric mean of all the others—requires transforming the data into different sets of pivot log-ratios and re-estimating the regression. By contrast, I show that a simple scaling factor recovers these ILR coefficients and their standard errors from a single ALR regression, obviating the need for multiple pivot permutations. Coenders and Pawlowsky-Glahn (2020) demonstrate that alternative log-ratios are different reparametrizations of the same model. Hence, the coefficients corresponding to the ALRs and ILRs both map directly into log-contrasts. In Sect. 6, I demonstrate that employing log-ratios does not exacerbate the multicollinearity that naturally arises in regressions that include compositional data. The results are illustrated throughout using cross-country economic growth regressions in which the explanatory variables associated with neoclassical growth theory are noncompositional, while additional variables—here religious shares—are compositional.²

2 Linear regression and compositional data: omitted components

Consider a generic linear model:

¹This is common in all the social sciences: labour economics (Burke and Sass 2013); development economics (Berman et al. 2017); political economy (Anagol and Fujiwara 2016); finance (Catherine et al. 2024); sociology (Logan et al. 2012; Hamm and McDonald 2015); and political science (Trounstine 2016).

²In Chen et al. (2017), models with only compositional covariates and real dependent variables are ‘Type 1’, all real covariates and compositional dependent variables ‘Type 2’, and models with compositional covariates and compositional dependent variables ‘Type 3’. The models here, with real dependent variables but a mixture of compositional and real covariates, represent a fourth type.

$$\mathbf{y} = \mathbf{N}\hat{\boldsymbol{\beta}}_N + \mathbf{C}\hat{\boldsymbol{\beta}}_C + \hat{\boldsymbol{\varepsilon}}, \quad (1)$$

where $\mathbf{y}$ is the observed dependent-variable vector, $\mathbf{N}$ is an $n \times K$ matrix of standard, noncompositional data and $\mathbf{C}$ is an $n \times (D + 1)$ matrix where the last column is the vector $\mathbf{1}_n$, that generates the intercept in the regression, and the other D columns represent compositional data constituting n points on the sample space $\mathbb{S}^D$, the $D - 1$ simplex:

$$\mathbb{S}^D = \left\{ [c_1, c_2, \dots, c_D]' \in \mathbb{R}_+^D \mid \sum_{i=1}^D c_i = \kappa \right\}. \quad (2)$$

The compositional covariates sum to a positive constant κ (often, though not always equal to 1), and so the columns of the partitioned covariate matrix $\mathbf{X} = (\mathbf{N}, \mathbf{C})$ are linearly dependent when an intercept is included. Consequently, $\mathbf{X}'\mathbf{X}$ is singular and the unconstrained coefficient vector in (1) is not uniquely identified. To ensure the covariate matrix has full column rank, the standard approach is to remove one compositional covariate, with its effect absorbed by the intercept.

Define the $(D + 1) \times D$ matrix:

$$\mathbf{Q}_{D,d} = \begin{bmatrix} \mathbf{I}_{d-1,d-1} & \mathbf{0}_{d-1,D-d+1} \\ \mathbf{0}_{1,d-1} & \mathbf{0}_{1,D-d+1} \\ \mathbf{0}_{D-d+1,d-1} & \mathbf{I}_{D-d+1,D-d+1} \end{bmatrix}.$$

Then the $n \times D$ matrix $\mathbf{C}_{\setminus d} = \mathbf{C}\mathbf{Q}_{D,d}$ is the matrix $\mathbf{C}$ with the $d \in \{1, \dots, D + 1\}$ column removed and $\mathbf{C}_{\setminus D+1} = \mathbf{C}\mathbf{Q}_{D,D+1}$ represents the raw data in the simplex (2).

Now the equation can be estimated using ordinary least squares (OLS):

$$\mathbf{y} = \mathbf{N}\hat{\boldsymbol{\beta}}_{N\setminus d} + \mathbf{C}_{\setminus d}\hat{\boldsymbol{\beta}}_{C\setminus d} + \hat{\boldsymbol{\varepsilon}}_{\setminus d}, \quad (3)$$

where $\hat{\boldsymbol{\beta}}_{C\setminus d}$ is the length D vector of estimated coefficients associated with the compositional data and the intercept term after the d^{th} column has been removed, and $\hat{\boldsymbol{\varepsilon}}_{\setminus d}$ is the vector of OLS residuals. The term $\hat{\boldsymbol{\beta}}_{N\setminus d}$ represents the length K vector of coefficients that correspond to the noncompositional data when the equation is estimated with the d^{th} column removed from $\mathbf{C}$. The case $d = D + 1$ returns the raw compositional data without the intercept column; the omitted-component regressions considered below remove one of the D compositional parts while retaining an intercept.

Define the projection matrices $\mathbf{P}_N \equiv \mathbf{N}(\mathbf{N}'\mathbf{N})^{-1}\mathbf{N}'$ and $\mathbf{P}_{C\setminus d} \equiv \mathbf{C}_{\setminus d}(\mathbf{C}_{\setminus d}'\mathbf{C}_{\setminus d})^{-1}\mathbf{C}_{\setminus d}'$. Furthermore define the $n \times (K + D)$ partitioned design matrix which contains both the variables in $\mathbf{N}$ and the vari-

ables in $C_{\setminus d}$: $X_{\setminus d} \equiv \begin{bmatrix} N \\ C_{\setminus d} \end{bmatrix}$ and the corresponding projection matrix $P_{X_{\setminus d}} \equiv X_{\setminus d} (X'_{\setminus d} X_{\setminus d})^{-1} X'_{\setminus d}$.

Using the standard partitioned-regression identities, equivalently the Frisch–Waugh–Lovell theorem, the solutions to the normal equations associated with (3) for the vector of estimated coefficients and the partitioned variance matrix (under homoscedasticity) are:

$$\begin{bmatrix} \hat{\beta}_{N_{\setminus d}} \\ \hat{\beta}_{C_{\setminus d}} \end{bmatrix} = \begin{bmatrix} (N' (I_n - P_{C_{\setminus d}}) N)^{-1} N' (I_n - P_{C_{\setminus d}}) \mathbf{y} \\ (C'_{\setminus d} (I_n - P_N) C_{\setminus d})^{-1} C'_{\setminus d} (I_n - P_N) \mathbf{y} \end{bmatrix} \quad (4)$$

$$\text{Var} \begin{bmatrix} \hat{\beta}_{N_{\setminus d}} \\ \hat{\beta}_{C_{\setminus d}} \end{bmatrix} = \frac{\hat{\epsilon}'_{\setminus d} \hat{\epsilon}_{\setminus d}}{n - K - D} \begin{bmatrix} (N' (I_n - P_{C_{\setminus d}}) N)^{-1} & - (N' N)^{-1} N' C_{\setminus d} (C'_{\setminus d} (I_n - P_N) C_{\setminus d})^{-1} \\ - (C'_{\setminus d} (I_n - P_N) C_{\setminus d})^{-1} C'_{\setminus d} N (N' N)^{-1} & (C'_{\setminus d} (I_n - P_N) C_{\setminus d})^{-1} \end{bmatrix} \quad (5)$$

where I_n is an $n \times n$ identity matrix and the vector of residuals is:

$$\hat{\epsilon}_{\setminus d} \equiv (I_n - P_{X_{\setminus d}}) \mathbf{y}. \quad (6)$$

In what follows, $\text{Var}(\hat{\beta}_{N_{\setminus d}})$ and $\text{Var}(\hat{\beta}_{C_{\setminus d}})$ denote the corresponding upper-left and lower-right diagonal blocks of the full covariance matrix in (5).

3 A lack of permutation invariance

Each permutation of (3), by which I mean the choice of the omitted component d , is a different representation of the same model. Most attributes are unchanged, but some do depend on the omitted component.

Let $e_1, \dots, e_D$ denote the standard basis vectors in $\mathbb{R}^D$, 1_{D-1} a column vector of ones of length $D - 1$ and define

$$\mathbf{v} \equiv \begin{bmatrix} -1_{D-1} \\ 1 \end{bmatrix}.$$

For $j = 1, \dots, D - 1$, define the adjacent-transposition permutation matrix $S_j \equiv [e_1, \dots, e_{j-1}, e_{j+1}, e_j, e_{j+2}, \dots, e_D]$ and for $i = 1, \dots, D - 1$, $A_i \equiv [e_1, \dots, e_{i-1}, \mathbf{v}, e_{i+1}, \dots, e_D]$. Define the $D \times D$ matrix for $d \neq f$:

$$L_{d,f} \equiv \begin{cases} A_d S_d S_{d+1} \cdots S_{f-2} & \text{for all } d \leq f-2 \\ A_d & \text{for all } d = f-1 \\ A_{d-1} & \text{for all } d = f+1 \\ A_{d-1} S_{d-2} S_{d-3} \cdots S_f & \text{for all } d \geq f+2 \end{cases} \quad (7)$$

Postmultiplying any $C_{\setminus d}$, $d \in \{1, \dots, D\}$ by A_i replaces the remaining i^{th} column in that matrix with the d^{th} column in the complete matrix C , and postmultiplying the result by the permutation matrix S_j exchanges the j and $j+1$ columns in the new matrix. Hence for all $d, f \in \{1, \dots, D\}$ where $d \neq f$: $C_{\setminus d} = C_{\setminus f} L_{d,f}$, $C_{\setminus f} = C_{\setminus d} L_{f,d}$ and $L_{d,f} = (L_{f,d})^{-1}$.

Lemma 1 *The projection matrices associated with the matrices C and X and error terms in (3) are permutation-invariant: $P_{C_{\setminus d}} = P_{C_{\setminus f}} = P_C$, $P_{X_{\setminus d}} = P_{X_{\setminus f}} = P_X$ and $\hat{\varepsilon}_{\setminus d} = \hat{\varepsilon}_{\setminus f} = \hat{\varepsilon}$ for all $d, f \in \{1, \dots, D\}$.*

Proof See Appendix A.1. The estimated coefficients associated with the noncompositional data N in (3), and the corresponding upper-left $K \times K$ block of the partitioned covariance matrix, are permutation-invariant. $\square$

Theorem 1 *The estimated coefficients associated with the noncompositional data N in (3), and the corresponding upper-left block $\text{Var}(\hat{\beta}_{N_{\setminus d}})$, of the partitioned variance-covariance matrix (5), are permutation-invariant: $\hat{\beta}_{N_{\setminus d}} = \hat{\beta}_{N_{\setminus f}} = \hat{\beta}_N$ and $\text{Var}(\hat{\beta}_{N_{\setminus d}}) = \text{Var}(\hat{\beta}_{N_{\setminus f}}) = \text{Var}(\hat{\beta}_N)$ for all $d, f \in \{1, \dots, D\}$. Furthermore, the R^2 and F -test for the regression are invariant to which compositional variable is omitted from the regression.*

Proof Follows directly from (4), the upper left-hand block in (5) and Lemma 1. $\square$

Theorem 2 *F -tests for the joint hypothesis that the first $D - 1$ elements of $\hat{\beta}_{C_{\setminus d}}$ are zero are invariant to the choice of $d \in \{1, \dots, D\}$.*

Proof See Appendix A.2. $\square$

Together, Theorems 1 and 2 imply that the choice of the omitted compositional variable does not affect the noncompositional coefficients, overall fit, or the joint explanatory power of the composition. This does not hold for the coefficients associated with the compositional data themselves, or the intercept.

Theorem 3 *The estimated coefficients associated with the compositional data C in (3), and the corresponding lower-right block $\text{Var}(\hat{\beta}_{C_{\setminus d}})$ of the partitioned vari-*

ance-covariance matrix (5), are generically not permutation-invariant. Specifically, $\hat{\beta}_{C \setminus d} = L_{f,d} \hat{\beta}_{C \setminus f}$ and $\text{Var}(\hat{\beta}_{C \setminus d}) = L_{f,d} \text{Var}(\hat{\beta}_{C \setminus f}) L'_{f,d}$.

Proof See Appendix A.3. $\square$

Define for $i < d$ [$i > d$], $\hat{\beta}_{C \setminus d}^i$ as element i [$i - 1$] in $\hat{\beta}_{C \setminus d}$, $\text{Var}(\hat{\beta}_{C \setminus d}^i)$ and $\text{Cov}(\hat{\beta}_{C \setminus d}^i, \hat{\beta}_{C \setminus d}^j)$ as the corresponding elements in the matrix $\text{Var}(\hat{\beta}_{C \setminus d})$. Theorem 3 implies that for any $i \neq d$:

$$\{\hat{\beta}_{C \setminus d}^i, \text{Var}(\hat{\beta}_{C \setminus d}^i)\} = \begin{cases} \{\hat{\beta}_{C \setminus f}^i - \hat{\beta}_{C \setminus f}^d, \text{Var}(\hat{\beta}_{C \setminus f}^i) + \text{Var}(\hat{\beta}_{C \setminus f}^d) - 2\text{Cov}(\hat{\beta}_{C \setminus f}^i, \hat{\beta}_{C \setminus f}^d)\}, & i \neq f, D, \\ \{-\hat{\beta}_{C \setminus f}^d, \text{Var}(\hat{\beta}_{C \setminus f}^d)\}, & i = f, \\ \{\hat{\beta}_{C \setminus f}^D + \hat{\beta}_{C \setminus f}^d, \text{Var}(\hat{\beta}_{C \setminus f}^D) + \text{Var}(\hat{\beta}_{C \setminus f}^d) + 2\text{Cov}(\hat{\beta}_{C \setminus f}^D, \hat{\beta}_{C \setminus f}^d)\}, & i = D. \end{cases} \quad (8)$$

Neither the estimated coefficients $\hat{\beta}_{C \setminus d}$ nor the estimated variances $\text{Var}(\hat{\beta}_{C \setminus d})$ corresponding to the different variables in C are permutation-invariant. As the different variables in C are not free to vary independently, each element in the vector of coefficients $\hat{\beta}_{C \setminus d}$ in (4) measures the effect of an increase in the value of the corresponding variable in $C \setminus d$ in relation to the excluded d^{th} variable.

Theorem 3 allows recovery of the coefficients from any alternative omitted category. However, the variances of the corresponding coefficients alone, i.e. the diagonal components of $\text{Var}(\hat{\beta}_{C \setminus d})$ —all that is usually reported in empirical research—are insufficient to derive the diagonal components of $\text{Var}(\hat{\beta}_{C \setminus f})$.

From (8) $\frac{\hat{\beta}_{C \setminus d}^i}{\sqrt{\text{Var}(\hat{\beta}_{C \setminus d}^i)}} = \frac{\hat{\beta}_{C \setminus f}^i - \hat{\beta}_{C \setminus f}^d}{\sqrt{\text{Var}(\hat{\beta}_{C \setminus f}^i) + \text{Var}(\hat{\beta}_{C \setminus f}^d) - 2\text{Cov}(\hat{\beta}_{C \setminus f}^i, \hat{\beta}_{C \setminus f}^d)}}$, need not equal $\frac{\hat{\beta}_{C \setminus f}^i}{\sqrt{\text{Var}(\hat{\beta}_{C \setminus f}^i)}}$ and the t -statistics for the same category i can differ across the different

permutations. Because the applied literature motivating this paper typically report conventional standard errors, t -statistics, and significance levels, I retain that focus here; permutation or bootstrap inference for log-contrast hypotheses is a natural extension. Researchers may be tempted to favour the permutation that appears most convincing, perhaps one that yields the most p -values that cross a desired threshold of significance or maximises the number of “stars”, in a manner analogous to p -hacking in Brodeur et al. (2016). To illustrate, I estimate the cross-country growth regression in Mankiw et al. (1992) using IMF data. The dependent variable y is the log change in GDP per capita (PPP) from 2001 to 2020. The level of log GDP per capita (PPP) in 2001, together with average savings and population growth rates over 2001–2020,

constitute the noncompositional covariates in N . Alongside these variables, which are associated with the canonical neoclassical (Solow) growth model, it is common to include additional covariates that capture cross-country heterogeneity and are often compositional.³ In the examples that follow, I consolidate the data for 35 religious affiliations from the World Religion Project (WRP) (Maoz and Henderson 2013) for 164 countries into $D = 10$ broad categories expressed as population shares, which, together with a column of ones, form matrix C .⁴

Each column of Table 1 reports a different permutation of (3), with one religious category omitted in turn. The results in the first three rows are consonant with the predictions of the neoclassical growth model: the coefficients on log GDP in levels and population growth are negative and the coefficient on savings is positive; all are statistically significant. In accordance with Theorem 1, as each successive religious category is excluded, the coefficient estimates and standard errors associated with the noncompositional data, as well as the values of R^2 , remain unchanged.

Column 1 of Table 1 excludes the Catholic share of the population and the coefficient on Orthodox Christians, relative to the excluded category, is not statistically significant. Column 2 reinstates Catholics but excludes Protestants, and here the coefficient on Orthodox Christians, relative to the excluded category is significant at the 5% level. Although each permutation is equally valid, the number of reported “stars” ranges from zero in columns 6 and 9, where Jews or Hindus are omitted, to fourteen in column 7, where Buddhists are omitted. One solution is to estimate all the parameters in (1) as a contrast model by constraining the first D elements of $\hat{\beta}_C$ to sum to zero. This resolves the problem of permutation invariance, but two issues remain.

First, suppose the link between economic growth and religious composition in Table 1 reflects not merely statistical correlation but a causal relationship. In 87 countries in our sample, the Buddhist share is zero, and in another 32 it is less than 0.1% of the population. Muslims constitute 99% or more of the population in three countries—Algeria, Tunisia, and Yemen—less than 0.1% in 22 countries, and zero in seven.⁵ A counterfactual that raises the Muslim share and lowers the Buddhist share by the same amount would push the former above 100% in some countries and the latter below zero.

To visualise this issue, I consolidate the ten religious categories into three broader groups—Christians, Muslims, and Other—and plot the amalgamated religious shares for each of the 164 countries on the ternary diagram in Fig. 1. Including raw shares in a regression means treating compositional data as real coordinates relative to a canonical basis. Yet applying Euclidean geometry to a simplex does not yield a vector space: ordinary operations such as adding two vectors as though they were Cartesian coordinates, or multiplying one vector by a scalar, can generate coordinates that fall

³ Age structure (Lindh 1999), shares of GDP or government spending (Devarajan et al. 1996; Voigt et al. 2015), languages spoken (Hall and Jones 1999; Rodrik et al. 2004), and religious affiliation (Hall and Jones 1999; Barro and McCleary 2003; Noland 2005; McCleary and Barro 2006).

⁴ See Figure B.2 in Online Appendix B.3 for the contribution biplot of the religion data.

⁵ In thirty countries there are no Hindus, Jews, or Orthodox Christians. See Online Appendix B.1 for complete lists.

Table 1 Omitted category regressions (3)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GDP (2001)	-0.167*** (0.022)	-0.167*** (0.022)	-0.167*** (0.022)	-0.167*** (0.022)	-0.167*** (0.022)	-0.167*** (0.022)	-0.167*** (0.022)	-0.167*** (0.022)	-0.167*** (0.022)	-0.167*** (0.022)
Pop. Growth	-0.957*** (0.150)	-0.957*** (0.150)	-0.957*** (0.150)	-0.957*** (0.150)	-0.957*** (0.150)	-0.957*** (0.150)	-0.957*** (0.150)	-0.957*** (0.150)	-0.957*** (0.150)	-0.957*** (0.150)
Savings	0.098** (0.042)	0.098** (0.042)	0.098** (0.042)	0.098** (0.042)	0.098** (0.042)	0.098** (0.042)	0.098** (0.042)	0.098** (0.042)	0.098** (0.042)	0.098** (0.042)
Catholic		0.175 (0.139)	-0.204 (0.135)	0.158 (0.359)	0.012 (0.091)	-0.378 (0.410)	-0.555*** (0.163)	0.657 (0.553)	-0.230 (0.229)	-0.139 (0.180)
Protestant	-0.175 (0.139)		-0.380** (0.163)	-0.017 (0.388)	-0.164 (0.126)	-0.554 (0.419)	-0.730*** (0.186)	0.482 (0.559)	-0.405 (0.249)	-0.315 (0.212)
Orthodox	0.204 (0.135)	0.380** (0.163)		0.363 (0.375)	0.216 (0.152)	-0.174 (0.425)	-0.350* (0.190)	0.861 (0.560)	-0.025 (0.250)	0.065 (0.198)
Other Christ	-0.158 (0.359)	0.017 (0.388)	-0.363 (0.375)		-0.147 (0.353)	-0.537 (0.534)	-0.713* (0.376)	0.499 (0.656)	-0.388 (0.421)	-0.298 (0.414)
Muslim	-0.012 (0.091)	0.164 (0.126)	-0.216 (0.152)	0.147 (0.353)		-0.390 (0.411)	-0.566*** (0.165)	0.645 (0.558)	-0.241 (0.233)	-0.151 (0.175)
Jewish	0.378 (0.410)	0.554 (0.419)	0.174 (0.425)	0.537 (0.534)	0.390 (0.411)		-0.176 (0.434)	1.035 (0.680)	0.148 (0.460)	0.239 (0.437)
Buddhist	0.555*** (0.163)	0.730*** (0.186)	0.350* (0.190)	0.713* (0.376)	0.566*** (0.165)	0.176 (0.434)		1.212** (0.607)	0.325 (0.275)	0.415* (0.231)
Other Eastern	-0.657 (0.553)	-0.482 (0.559)	-0.861 (0.560)	-0.499 (0.656)	-0.645 (0.558)	-1.035 (0.680)	-1.212** (0.607)		-0.887 (0.590)	-0.796 (0.580)
Hindu	0.230 (0.229)	0.405 (0.249)	0.025 (0.250)	0.388 (0.421)	0.241 (0.233)	-0.148 (0.460)	0.607 (0.275)	0.887 (0.590)		0.090 (0.263)
Other	0.139 (0.180)	0.315 (0.212)	-0.065 (0.198)	0.298 (0.414)	0.151 (0.175)	-0.239 (0.437)	-0.415* (0.231)	0.796 (0.580)	-0.090 (0.263)	
Constant	2.053*** (0.204)	1.877*** (0.226)	2.257*** (0.208)	1.894*** (0.398)	2.041*** (0.204)	2.431*** (0.465)	2.607*** (0.233)	1.396*** (0.606)	2.282*** (0.296)	2.192*** (0.240)

Table 1 (continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Observations	164	164	164	164	164	164	164	164	164	164
R^2	0.509	0.509	0.509	0.509	0.509	0.509	0.509	0.509	0.509	0.509
Adjusted R^2	0.469	0.469	0.469	0.469	0.469	0.469	0.469	0.469	0.469	0.469
Mean VIF	1.34	1.82	1.91	8.29	1.31	10.54	2.36	18.73	3.86	2.64
Cond. Number	30.23	30.90	30.92	60.21	30.51	67.37	36.36	91.39	39.62	34.81

Dependent variable is log difference of per-capita GDP (PPP) between 2001 and 2020

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

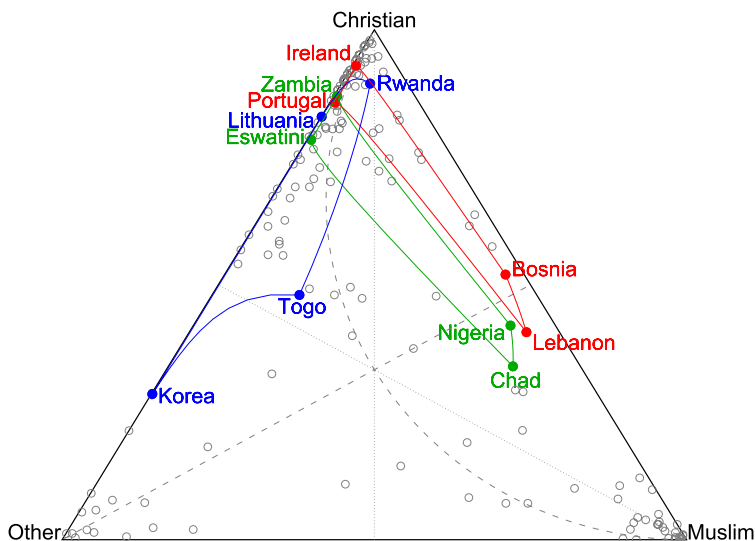


Fig. 1 Ternary diagram for three-part religious composition for the 164 countries in the dataset. Dashed curves represent the basis for ILR coordinates and dotted lines the basis for ALR coordinates. Curves in blue, green and red represent the sides of rectangles in terms of Aitchison geometry. See Table B.1 in the online appendix for Euclidean and Aitchison distances between the vertices of the three rectangles

outside the simplex.⁶ A contrast model compounds this problem because any counterfactual experiment that shifts one share necessarily requires a reallocation from one or more of the remaining shares.⁷

Second, how informative are the Euclidean distances between compositional coordinates? In our dataset, the Muslim share of Italy's population is just over 1%, whereas in France it is 8%. That eightfold difference reflects France's much longer and broader colonial presence in the Middle East and Africa. But the Muslim share

⁶To overcome these limitations, the concepts of Aitchison (1986) clo-

$$\text{sure } \mathcal{C}(z) = \left[\frac{z_1}{\sum_{i=1}^D z_i}, \frac{z_2}{\sum_{i=1}^D z_i}, \dots, \frac{z_D}{\sum_{i=1}^D z_i} \right], \text{perturbation } \mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \dots, x_D y_D)$$

in place of vector addition, and powering $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha)$ in place of scalar multiplication, are sufficient to make the simplex a vector space (the point where all shares are equal is the barycentre), (Billheimer et al. 2001; Pawłowsky-Glahn and Egozcue 2001). Adding the Aitchison inner product and norm, the simplex becomes a Hilbert space. See Online Appendix B.2.

$$\mathcal{C}(z) = \left[\frac{z_1}{\sum_{i=1}^D z_i}, \frac{z_2}{\sum_{i=1}^D z_i}, \dots, \frac{z_D}{\sum_{i=1}^D z_i} \right], \text{perturbation}$$

$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \dots, x_D y_D)$ in place of vector addition, and powering

$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha)$ in place of scalar multiplication, are sufficient to make the simplex a vector space (the point where all shares are equal is the barycentre), (Billheimer et al. 2001; Pawłowsky-Glahn and Egozcue 2001). Adding the Aitchison inner product and norm, the simplex becomes a Hilbert space. See Online Appendix B.2.

⁷See Figure B.3 in Online Appendix B.3.

of the population is 90% in Oman and 97% in Jordan: is the same 7-percentage-point difference equally meaningful? In recent decades, the Protestant share has also grown rapidly in some historically Catholic Latin American countries: Protestants now constitute 27% of the population in Brazil, 15% in Colombia, and only 2% in Ecuador. In Denmark and Norway, they constitute 81% of the population, and in Sweden 68%—the same percentage-point gap as between Brazil and Colombia, and nearly the same as between Colombia and Ecuador. It is hard to regard these differences as equivalent: the Protestant share of Brazil's population is more than thirteen times Ecuador's. Yet the coefficients in Table 1 are estimated as though they were.

4 Additive logarithmic ratios

A way forward is to apply logarithmic transformations that map points in the simplex $\mathbb{S}^D$ in (2) into a subset of Euclidean space $\mathbb{R}^D$. First, we must replace any zeros in $\text{CQ}_{D,D+1}$ with small positive values and adjust the remaining data so that it sums to one.⁸ Let $\check{C} = \left[\log \left(c_i^{(\ell)} \right)_{\ell=1, \dots, n; i=1, \dots, D} \vdots 1_n \right]$ denote the matrix formed by taking natural logarithms of the compositional data $\text{CQ}_{D,D+1}$ and appending a column of ones. Because the logarithmic transformation is nonlinear, we can regress $\mathbf{y}$ on $\mathbf{N}$ and $\check{C}$:

$$\mathbf{y} = \mathbf{N}\check{\beta}_N + \check{C}\check{\beta}_C + \check{\epsilon}. \quad (9)$$

The first D elements of $\check{\beta}_C$ represent elasticities rather than coefficients on raw shares. As all categories are included, permutation invariance is no longer an issue. Constraining the first D elements of $\check{\beta}_C$ to sum to zero yields a log-contrast model, presented (without $\mathbf{N}$) in Aitchison and Bacon-Shone (1984).

Using log-contrasts rather than contrasts on the original scale has two benefits. First, although in this example the religion composition includes an absorbing “other” category and is exhaustive, a log-contrast formulation remains valid and interpretable even when the observed parts form only a subcomposition of a larger whole—that is, when unobserved categories exist or are excluded from the model.

Second, unlike contrast models on the original scale, the log-contrast model in (9) permits counterfactual experiments without producing invalid compositions outside the simplex, provided changes in one share are offset by reallocation. A less cumbersome alternative, especially for very large datasets, is to use additive log-ratios (ALRs), again following Aitchison and Bacon-Shone (1984), which avoid reallocation altogether.

Let $0_{m,n}$ be an $m \times n$ matrix of zeros. Define the $D \times (D - 1)$ and $(D + 1) \times D$ matrices:

⁸Following Aitchison (1986), p. 269, I set the Z zero components for each observation equal to $\delta(Z + 1)(D - Z)/D^2$ and reduce the positive elements by $\delta Z(Z + 1)/D^2$. Setting $\delta = 10^{-5}$ replaces the zeros by shares of between 1 and 3 per million people.

$$\mathbf{F}_d = \begin{bmatrix} \mathbf{I}_{d-1,d-1} & \mathbf{0}_{d-1,D-d} \\ \mathbf{0}_{D-d,d-1} & \mathbf{I}_{D-d,D-d} \end{bmatrix}, \quad \tilde{\mathbf{F}}_d = \begin{bmatrix} \mathbf{F}_d & \mathbf{0}_{D,1} \\ \mathbf{0}_{1,D-1} & 1 \end{bmatrix}, \quad (10)$$

Postmultiplying $\tilde{\mathbf{C}}$ by $\tilde{\mathbf{F}}_d$ generates an $n \times D$ matrix $\tilde{\mathbf{C}}_{/d} = \tilde{\mathbf{C}}\tilde{\mathbf{F}}_d$, where each column is the log deviation of the remaining $i \neq d$ columns from the d^{th} variable, and the last column is an $n \times 1$ vector of ones.⁹ The regression with ALR coordinates (Aitchison and Bacon-Shone 1984), augmented to include noncompositional covariates is

$$\mathbf{y} = \mathbf{N}\tilde{\boldsymbol{\beta}}_{N/d} + \tilde{\mathbf{C}}_{/d}\tilde{\boldsymbol{\beta}}_{C/d} + \tilde{\boldsymbol{\varepsilon}}_{/d}, \quad d \in \{1, \dots, D\}, \quad (11)$$

and changes to any or all $D - 1$ log differences in (11) generate compositions that are always within the simplex.

Intuitively, given the close relationship between the log-contrast model and ALR, the fitted model should not depend on which category is used as the ALR baseline. Define the $D \times D$ matrix $\mathbf{M}_{d,f}$ that maps the ALR representation $\tilde{\mathbf{C}}_{/f}$, in which category f is the baseline and all other categories are expressed relative to it, into the corresponding representation $\tilde{\mathbf{C}}_{/d}$, in which category d is the baseline:

$$\mathbf{M}_{d,f} \equiv \begin{cases} \mathbf{H}_d \mathbf{S}_d \mathbf{S}_{d+1} \cdots \mathbf{S}_{f-2} & \text{for all } d \leq f-2 \\ \mathbf{H}_d & \text{for all } d = f-1 \\ \mathbf{H}_{d-1} & \text{for all } d = f+1 \\ \mathbf{H}_{d-1} \mathbf{S}_{d-2} \mathbf{S}_{d-3} \cdots \mathbf{S}_f & \text{for all } d \geq f+2 \end{cases} \quad (12)$$

$$\text{where: } \mathbf{H}_j \equiv \begin{bmatrix} \mathbf{I}_{j-1,j-1} & \mathbf{0}_{j-1,D-j+1} \\ -\mathbf{1}'_{D-1} & 0 \\ \mathbf{0}_{D-j,j} & \mathbf{I}_{D-j,D-j} \end{bmatrix}.$$

Postmultiplying any $\tilde{\mathbf{C}}_{/d}$ by $\mathbf{H}_j$ transforms the log deviations with respect to variable d into log deviations with respect to variable j , and postmultiplying the result by the permutation matrix $\mathbf{S}_j$ exchanges the j and $j+1$ columns in the new matrix. Hence for all $d, f \in \{1, \dots, D\}$ where $d \neq f$: $\tilde{\mathbf{C}}_{/d} = \tilde{\mathbf{C}}_{/f} \mathbf{M}_{d,f}$, $\tilde{\mathbf{C}}_{/f} = \tilde{\mathbf{C}}_{/d} \mathbf{M}_{f,d}$ and $\mathbf{M}_{d,f} = (\mathbf{M}_{f,d})^{-1}$. Define two projection matrices $\tilde{\mathbf{P}}_{C/d} \equiv \tilde{\mathbf{C}}_{/d} (\tilde{\mathbf{C}}'_{/d} \tilde{\mathbf{C}}_{/d})^{-1} \tilde{\mathbf{C}}'_{/d}$ and $\tilde{\mathbf{P}}_{X/d} \equiv \tilde{\mathbf{X}}_{/d} (\tilde{\mathbf{X}}'_{/d} \tilde{\mathbf{X}}_{/d})^{-1} \tilde{\mathbf{X}}'_{/d}$ where $\tilde{\mathbf{X}}_{/d} \equiv [\mathbf{N}; \tilde{\mathbf{C}}_{/d}]$.

Lemma 2 *The projection matrices $\tilde{\mathbf{P}}_{C/d}$ and $\tilde{\mathbf{P}}_{X/d}$, as well as the error terms $\tilde{\boldsymbol{\varepsilon}}_{/d}$ in (11) are all permutation-invariant. Hence $\tilde{\mathbf{P}}_{C/d} = \tilde{\mathbf{P}}_{C/f} = \tilde{\mathbf{P}}_C$, $\tilde{\mathbf{P}}_{X/d} = \tilde{\mathbf{P}}_{X/f} = \tilde{\mathbf{P}}_X$ and $\tilde{\boldsymbol{\varepsilon}}_{/d} = \tilde{\boldsymbol{\varepsilon}}_{/f} = \tilde{\boldsymbol{\varepsilon}}$ for all $d, f \in \{1, \dots, D\}$.*

⁹ In Sect. 3 we use the subscript $\backslash d$ to designate the d^{th} variable or column omitted from the data matrix and its associated coefficients, and in Sects. 4 and 5 we use the subscript $/d$ to designate log deviations from the d^{th} variable.

Proof See Appendix A.4. $\square$

The estimated coefficients in (11) can be written as:

$$\begin{bmatrix} \tilde{\beta}_N \\ \tilde{\beta}_{C/d} \end{bmatrix} = \begin{bmatrix} (N' (I_n - \tilde{P}_C) N)^{-1} N' (I_n - \tilde{P}_C) \mathbf{y} \\ (\tilde{C}'_d (I_n - P_N) \tilde{C}_d)^{-1} \tilde{C}'_d (I_n - P_N) \mathbf{y} \end{bmatrix} \quad (13)$$

and the estimated partitioned variance matrix is:

$$\text{Var} \begin{bmatrix} \tilde{\beta}_N \\ \tilde{\beta}_{C/d} \end{bmatrix} = \frac{\tilde{\epsilon}' \tilde{\epsilon}}{n - K - D} \begin{bmatrix} (N' (I_n - \tilde{P}_C) N)^{-1} & - (N' N)^{-1} N' \tilde{C}_d (\tilde{C}'_d (I_n - P_N) \tilde{C}_d)^{-1} \\ - (\tilde{C}'_d (I_n - P_N) \tilde{C}_d)^{-1} \tilde{C}'_d N (N' N)^{-1} & (\tilde{C}'_d (I_n - P_N) \tilde{C}_d)^{-1} \end{bmatrix} \quad (14)$$

Denote $\text{Var}(\tilde{\beta}_{N/d})$ and $\text{Var}(\tilde{\beta}_{C/d})$ as the corresponding upper-left and lower-right diagonal blocks of the full covariance matrix in (14). From Lemma 2 we generate the log-ratio analogues to Theorems 1–2.

Theorem 4 *The estimated coefficients and variances associated with the noncompositional data N are permutation-invariant: $\tilde{\beta}_{N/d} = \tilde{\beta}_{N/f} = \tilde{\beta}_N$, $\text{Var}(\tilde{\beta}_{N/d}) = \text{Var}(\tilde{\beta}_{N/f}) = \text{Var}(\tilde{\beta}_N)$ for all $d, f \in \{1, \dots, D\}$ as is the R^2 for the regression (11). The F -test for the joint hypothesis that for any choice $d \in \{1, \dots, D\}$, the vector $\tilde{\beta}_{C/d} = 0$ is also permutation-invariant.*

Proof Follows directly from Lemma 2 and arguments in proofs of Theorems 1–2. $\square$

Theorem 5 *The estimated coefficients and variances associated with the compositional data that are common to both $\tilde{C}_d$ and $\tilde{C}_f$ in (11) are identical.*

Proof Follows the same logic as the proof in Theorem 3 and $M_{d,f}^{-1} = M_{f,d}$ since $\tilde{\beta}_{C/f} = M_{d,f} \tilde{\beta}_{C/d}$ and $\text{Var}(\tilde{\beta}_{C/f}) = M_{d,f} \text{Var}(\tilde{\beta}_{C/d}) M'_{d,f}$. The salient factor in $M_{d,f}$ is the matrix H_d (or H_{d-1}), analogous to A_d in Sect. 3. Unlike A_d , H_d is an identity matrix for all but row d (or $d - 1$). $\square$

So, in contrast to Theorem 3, not only are $\tilde{\beta}_N$ and its associated variance permutation-invariant, but so are all the coefficients that correspond to the compositional variables, along with the intercept.

The values of the coefficients (including the constant term) in $\tilde{\beta}_{C/d}$ and associated variances across the different permutations can be summarised by:

$$\left\{ \tilde{\beta}_{C/d}^i, \text{Var} \left(\tilde{\beta}_{C/d}^i \right) \right\} = \begin{cases} \left\{ \tilde{\beta}_{C/f}^i, \text{Var} \left(\tilde{\beta}_{C/f}^i \right) \right\}, i \neq f \text{ or } i = D \\ \left\{ - \sum_{\forall j \neq i, j \neq D} \tilde{\beta}_{C/f}^j, \sum_{\forall j \neq i, j \neq D} \sum_{\forall k \neq i, k \neq D} \text{Cov} \left(\tilde{\beta}_{C/f}^j, \tilde{\beta}_{C/f}^k \right) \right\}, i = f, i \neq D \end{cases} \quad (15)$$

In each column of Table 2, which represents a different permutation of (11), the missing religious category is no longer an excluded share, but one that serves as the denominator in the log-ratios.¹⁰ As the different permutations all map into the same log-contrast model, the common elements between $\tilde{\beta}_{C/d}$ and $\tilde{\beta}_{C/f}$, $f \neq d$, in Table 2 are identical, and as implied by (15), the f^{th} element of $\tilde{\beta}_{C/d}$ equals minus the sum of the remaining coefficients, excluding the intercept. The hypothetical sum of the coefficients associated with all D possible elements of $\check{C}$ is, by construction, equal to zero, and the constant term is permutation-invariant. Furthermore, all the common diagonal terms between $\text{Var}(\tilde{\beta}_{C/f})$ and $\text{Var}(\tilde{\beta}_{C/d})$ are identical. For any $f \neq d$, the variance of the coefficient on variable d , which is omitted in the d -based parametrization, can be recovered from the f -based parametrisation as the sum of all elements of $\text{Var}(\tilde{\beta}_{C/f})$.

5 Isometric log-ratios

Additive log-ratios (ALRs) are the logarithmic differences between any $D - 1$ of the first D components in C and the missing d^{th} component c_d :

$$\begin{aligned} & [\check{c}_1 - \check{c}_d, \check{c}_2 - \check{c}_d, \dots, \check{c}_{d-1} - \check{c}_d, \check{c}_{d+1} - \check{c}_d, \dots, \check{c}_D - \check{c}_d] \\ & = [\log(c_1 \oslash c_d), \log(c_2 \oslash c_d), \dots, \log(c_{d-1} \oslash c_d), \\ & \quad \log(c_{d+1} \oslash c_d), \dots, \log(c_D \oslash c_d)] \end{aligned} \quad (16)$$

where $\oslash$ represents Hadamard division and the logarithm is taken for each element in the vector, translating the points in the simplex $\mathbb{S}^D$ in (2) to the Euclidean space $\mathbb{R}^{D-1}$.

¹⁰Applying log-ratios to the compositional data does not change the behaviour of the noncompositional parts of the model (see plots of orthogonal components in Online Appendix B.3 Figures B.4-B.6) or the errors in a predictable way (see Figure B.7). Nor does it necessarily yield better-fitting regressions. Table 2 shows a slight improvement in $R^2(0.526)$ compared to Table 1 (0.509). However, if we consolidate Orthodox Christians and Other Christians into one category, but separate Muslims into two categories, Sunni Muslims and non Sunni Muslims, R^2 is 0.500 with additive log-ratios in Online Appendix B.3, Table B.3, and 0.509 in Table B.2, where compositional data omits one variable.

$R^2(0.526)$ compared to Table 1 (0.509). However, if we consolidate Orthodox Christians and Other Christians into one category, but separate Muslims into two categories, Sunni Muslims and non Sunni Muslims, R^2 is 0.500 with additive log-ratios in Online Appendix B.3, Table B.3, and 0.509 in Table B.2, where compositional data omits one variable.

Taking the natural exponentials of each component in the canonical basis and normalising so that each vector sums to one (applying the closure operator $\mathcal{C}$ as in Appendix A.2), generates the $D \times D$ matrix W whose D columns $w_1, \dots, w_D$ are the elements:

$$w_{i,j} = \begin{cases} \frac{e}{e + D - 1}, & \text{if } i = j, \\ \frac{1}{e + D - 1}, & \text{if } i \neq j. \end{cases} \quad (17)$$

In terms of Aitchison geometry, the matrix of compositional data can be expressed as:

$$\begin{aligned} CQ_{D,D+1} &= [(\check{c}_1^{(\ell)} \odot w_1) \oplus (\check{c}_2^{(\ell)} \odot w_2) \oplus \dots \oplus (\check{c}_D^{(\ell)} \odot w_D)]_{\ell=1}^n \\ &= [((\check{c}_1^{(\ell)} - \check{c}_d^{(\ell)}) \odot w_1) \oplus ((\check{c}_2^{(\ell)} - \check{c}_d^{(\ell)}) \odot w_2) \\ &\quad \oplus \dots \oplus ((\check{c}_{d-1}^{(\ell)} - \check{c}_d^{(\ell)}) \odot w_{d-1}) \oplus ((\check{c}_{d+1}^{(\ell)} - \check{c}_d^{(\ell)}) \odot w_{d+1}) \\ &\quad \oplus \dots \oplus ((\check{c}_D^{(\ell)} - \check{c}_d^{(\ell)}) \odot w_D)]_{\ell=1}^n, \end{aligned} \quad (18)$$

where $\oplus$ represents perturbation and $\odot$ powering (as defined in footnote 6 and Online Appendix B.2)—the equivalent in Aitchison geometry of addition and scalar multiplication respectively. The second equality means that the set of $D - 1$ vectors $[w_1, w_2, \dots, w_{d-1}, w_{d+1}, \dots, w_D]$ are a basis for the simplex upon which the additive log-ratios (16) are coordinates.

These basis vectors are neither normalised nor orthogonal. The Aitchison norm is $\|w_i\|_a = \sqrt{\frac{D-1}{D}}$, and the Aitchison inner product between any two i, j vectors is $\langle w_i, w_j \rangle_a = -1/D$. Hence the angle between the vectors (in radians) is $\arccos(\frac{1}{1-D})$ and the ALR transformation from the simplex $\mathbb{S}^D$ to Euclidean space $\mathbb{R}^{D-1}$ generates coordinates with respect to an oblique basis, though one that is asymptotically orthogonal as D increases. In Fig. 1 where $D=3$, the angles between the basis vectors (dotted lines) are equal to $\frac{2\pi}{3}$ or 120° and ALR translates the rectangles into parallelograms in Fig. 2a.

Alternatively, taking the Aitchison inner product between the points on the simplex and any orthonormal basis e on the simplex, transforms the data into a length $D - 1$ vector of isometric log-ratios (ILR) (Egozcue et al. 2003). This is equivalent to postmultiplying $\check{C}Q_{D,D+1}$ by the transpose of a $(D - 1) \times D$ contrast matrix U , where the $D - 1$ rows are centered log-ratios (CLRs, log deviations of each element from the geometric mean) of the chosen orthonormal basis e . Transposing U , the $D - 1$ columns in U' form a basis in a $D - 1$ subspace where for each basis vector, the sum of its D elements is zero. Postmultiplying an ALR composition by F^+ (the Moore-Penrose generalised inverse of F from (10)) generates a CLR, and postmultiplying that by U' generates an ILR composition.¹¹ The CLR is the unique symmetric log-ratio representation of the same underlying log-contrast, while the ILR coordi-

¹¹ Proposition 4 in (Egozcue et al. 2003).

Table 2 Additive log ratio regressions (11)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GDP (2001)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)
Pop. Growth	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)
Savings	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)
Catholic		-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)
Protestant	-0.002 (0.010)		-0.002 (0.010)	-0.002 (0.010)	-0.002 (0.010)	-0.002 (0.010)	-0.002 (0.010)	-0.002 (0.010)	-0.002 (0.010)	-0.002 (0.010)
Orthodox	0.020*** (0.006)	0.020*** (0.006)	0.020*** (0.006)	0.020*** (0.006)	0.020*** (0.006)	0.020*** (0.006)	0.020*** (0.006)	0.020*** (0.006)	0.020*** (0.006)	0.020*** (0.006)
Other Christ	0.001 (0.006)	0.001 (0.006)	0.001 (0.006)		0.001 (0.006)	0.001 (0.006)	0.001 (0.006)	0.001 (0.006)	0.001 (0.006)	0.001 (0.006)
Muslim	-0.009 (0.008)	-0.009 (0.008)	-0.009 (0.008)	-0.009 (0.008)		-0.009 (0.008)	-0.009 (0.008)	-0.009 (0.008)	-0.009 (0.008)	-0.009 (0.008)
Jewish	0.002 (0.010)	0.002 (0.010)	0.002 (0.010)	0.002 (0.010)	0.002 (0.010)		0.002 (0.010)	0.002 (0.010)	0.002 (0.010)	0.002 (0.010)
Buddhist	0.027*** (0.007)	0.027*** (0.007)	0.027*** (0.007)	0.027*** (0.007)	0.027*** (0.007)	0.027*** (0.007)	0.027*** (0.007)	0.027*** (0.007)	0.027*** (0.007)	0.027*** (0.007)
Other Eastern	-0.002 (0.011)	-0.002 (0.011)	-0.002 (0.011)	-0.002 (0.011)	-0.002 (0.011)	-0.002 (0.011)	-0.002 (0.011)	-0.002 (0.011)	-0.002 (0.011)	-0.002 (0.011)
Hindu	-0.007 (0.007)	-0.007 (0.007)	-0.007 (0.007)	-0.007 (0.007)	-0.007 (0.007)	-0.007 (0.007)	-0.007 (0.007)	-0.007 (0.007)		-0.007 (0.007)
Other	-0.027* (0.016)	-0.027* (0.016)	-0.027* (0.016)	-0.027* (0.016)	-0.027* (0.016)	-0.027* (0.016)	-0.027* (0.016)	-0.027* (0.016)	-0.027* (0.016)	-0.027* (0.016)
Constant	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)

Table 2 (continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Observations	164	164	164	164	164	164	164	164	164	164
R^2	0.526	0.526	0.526	0.526	0.526	0.526	0.526	0.526	0.526	0.526
Adjusted R^2	0.488	0.488	0.488	0.488	0.488	0.488	0.488	0.488	0.488	0.488
Mean VIF	2.73	2.65	4.45	3.84	3.44	2.53	3.30	1.98	3.39	1.61
Cond. Number	69.27	68.37	92.09	78.24	74.60	61.72	72.70	45.35	72.19	38.16

Dependent variable is log difference of per-capita GDP (PPP) between 2001 and 2020

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

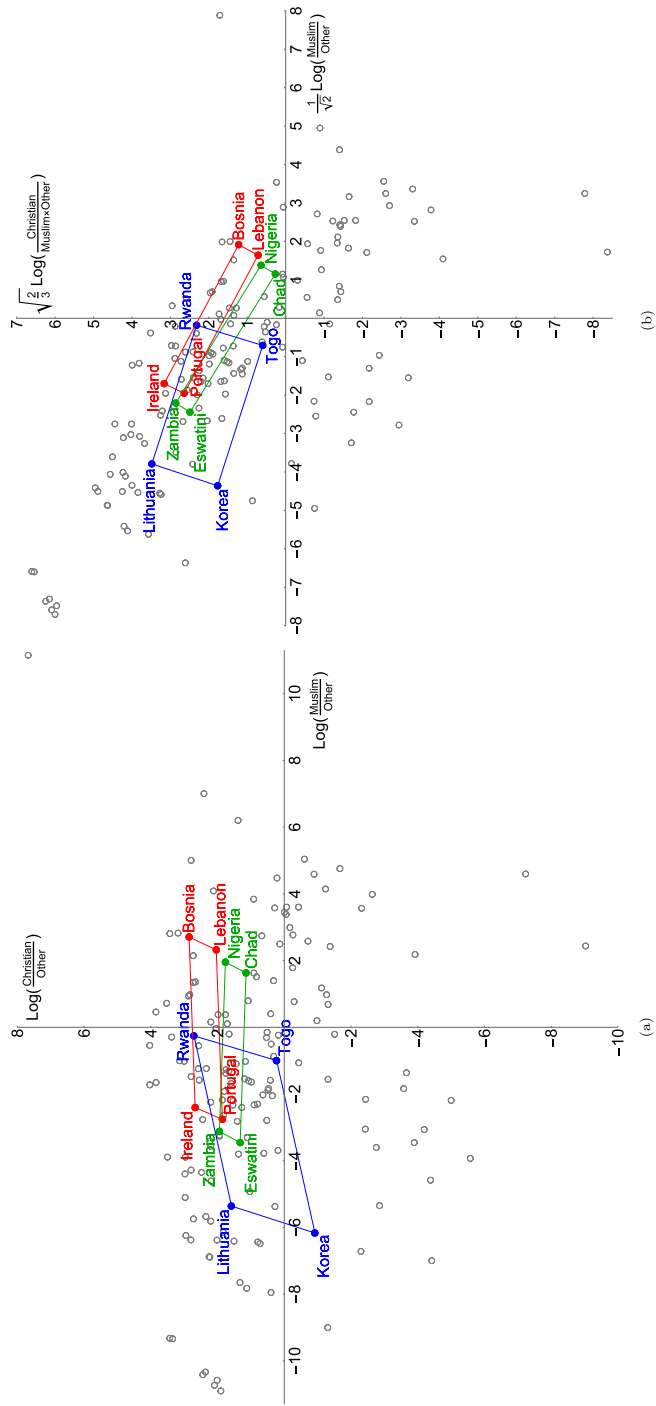


Fig. 2 (a) ALR transformation for the three-part composition in Fig. 1 represented in ALR axes. (b) ILR transformation for the three-part composition in Fig. 1 represented in ILR axes

nates connect that log-contrast to orthonormal parametrizations commonly used in regression applications.

Accommodating the vector of ones that generate the intercept term, we convert additive log-ratios to a generic set of isometric log-ratios by postmultiplying $\tilde{C}_{/d}$ by the $D \times D$ matrix $\bar{F}_d \bar{U}'$, where:

$$\bar{F}_d = \begin{bmatrix} - & - & \mathbf{F}_d^+ & - & \dots & - & - \\ & 1/D & 1/D & \dots & 1/D & - & - \end{bmatrix}, \quad (19)$$

and $\bar{U}$ is a generic contrast matrix U , appended with a row vector of ones, 1_D . Note that each coordinate in $\tilde{C}_{/d} \bar{F}_d$ is the sum of the CLR and the constant term $1/D$.

The number of potential orthonormal bases is vast. Pawlowsky-Glahn et al. (2015) suggest using the sequential binary partition method:

$$U_D = \begin{bmatrix} -\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} & 0 & \dots & \dots & 0 \\ -\frac{1}{2}\sqrt{\frac{2}{3}} & -\frac{1}{2}\sqrt{\frac{2}{3}} & \sqrt{\frac{2}{3}} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \dots & \sqrt{\frac{D-1}{D}} \end{bmatrix} \quad (20)$$

which generates a set of balances with these coordinates:

$$\bar{c}_{i/1}^{(\ell)} = \sqrt{\frac{i-1}{i}} \ln \left[c_i^{(\ell)} / \left(\prod_{j=1}^{i-1} c_j^{(\ell)} \right)^{\frac{1}{i-1}} \right], i = 2, \dots, D; \ell = 1, \dots, n, \quad (21)$$

also known as pivot log-ratios.¹² Equation (21) gives the coordinates associated with the pivot basis U_D : for each observation l , the first coordinate is the log deviation of $c_2^{(\ell)}$ from $c_1^{(\ell)}$, the second is the log deviation of $c_3^{(\ell)}$ from the geometric mean of $c_1^{(\ell)}$ and $c_2^{(\ell)}$, and the final coordinate is the log deviation of $c_D^{(\ell)}$ from the geometric mean of the preceding $D-1$ components. The last remaining term is potentially the most interesting as it represents the log deviation of $c_D^{(\ell)}$ from the geometric mean of all the preceding $D-1$ variables. Comparing Fig. 2a, b illustrates the change in coordinates induced by moving from ALR to ILR.

We can extend (20) to the other $h \in \{1, \dots, D-1\}$ by defining $U_h \equiv U_D S_h S_{h+1} \dots S_{D-1}$, which moves component h to the final position while preserving the order of the remaining $D-1$ components, so that the last pivot coordinate isolates component h . We then define $G_{d,h} \equiv \bar{F}_d \bar{U}_h'$, where the augmented matrix $\bar{U}_h$ is U_h with an additional final row vector of ones. The symmetry of the CLR means that $\tilde{C}_{/d} \bar{F}_d = \tilde{C}_{/f} \bar{F}_f$ for all $d, f \in \{1, \dots, D\}$. Hence, we can generate

¹²See Greenacre (2018)) for an alternative matrix transformation that generates pivot log-ratios.

a matrix of isometric log-ratios $\bar{C}_{/h}$, associated with $\bar{U}_h$, from any set of additive log-ratios so that $\bar{C}_{/h} = \bar{C}_{/d}G_{d,h} = \bar{C}_{/f}G_{f,h}$ and estimate any $h = 1, \dots, D$ different versions of the regression¹³

$$\mathbf{y} = N\bar{\beta}_{N/h} + \bar{C}_{/h}\bar{\beta}_{C/h} + \bar{\varepsilon}_{/h}. \quad (22)$$

Lemma 3 Define the matrix $\bar{X}_{/h} \equiv [N; \bar{C}_{/h}]$. The projection matrices associated with the ILR coordinates are permutation-invariant and equal to those associated with the ALR coordinates so that $\bar{P}_{C/h} = \bar{P}_C$ and $\bar{P}_{X/h} = \bar{P}_X$ for all $h \in \{1, \dots, D\}$.

Proof See Appendix A.5 □

Theorem 6 The estimated coefficients, variances, and residuals associated with the noncompositional data N in (22) are all permutation-invariant and identical to those in the regression using ALR coordinates (11): $\bar{\beta}_N = \hat{\beta}_N$, $\text{Var}(\bar{\beta}_N) = \text{Var}(\hat{\beta}_N)$ and $\bar{\varepsilon} = \hat{\varepsilon}$. The R^2 for (22) and the F -test for the joint hypothesis for the compositional variables $\bar{\beta}_{C/h} = 0$ are also identical to those in (11).

Proof Follows directly from Lemma 3. □

It follows that we can either estimate (22) or estimate (11) and use $\bar{\beta}_{C/h} = G_{d,h}^{-1}\tilde{\beta}_{C/d}$ and $\text{Var}(\bar{\beta}_{C/h}) = G_{d,h}^{-1}\text{Var}(\tilde{\beta}_{C/d})(G'_{d,h})^{-1}$ to generate coefficients and standard errors. Hron et al. (2012) suggest estimating all D permutations of (22) as in Table 3 to derive coefficients of particular interest—those associated with the log difference between one component and the geometric mean of all the others.

In column (1), the coefficient on the second category, Protestants, relates to the log deviation between their share and the share of Catholics, the first category. The coefficient on the third category, Orthodox Christians, relates to the log deviation between their share and the geometric mean of the shares of Protestants and Catholics. The final religious category, Other Religions, highlighted in bold, relates to the geometric mean of the shares of all other categories and is therefore the coefficient highlighted by Hron et al. (2012) as being of greatest interest.

In column (2), the coefficient for Hindus, highlighted in bold, corresponds to the log deviation between that share and the geometric mean of all the others. The remaining coefficients follow the same pattern as in column (1), except that Protestant, rather than Catholic, now appears in every geometric mean. In each subsequent column, the coefficients corresponding to the log deviation between that category and all nine remaining categories appear in ascending order. The omitted category on the diagonal is the category that appears in every geometric mean.

¹³ Each row of $\bar{C}_{/d}$ represents the $D - 1$ log deviations of the shares $i \in \{1, \dots, D\} : \{d\}$ with respect to the d^{th} share, and each row of $\bar{C}_{/h}$ represents the log deviations of the shares $i \in \{1, \dots, D\} \setminus \{h\}$ with respect to the cascading geometric means. If $h = d$, the first column in $\bar{C}_{/d}$ always equals the first column in $\bar{C}_{/d}$, divided by $\sqrt{2}$.

Table 3 Isometric log ratio regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GDP (2001)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)	-0.217*** (0.023)
Pop. Growth	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)	-0.971*** (0.143)
Savings	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)	0.126*** (0.041)
Catholic		-0.001 (0.012)	-0.016** (0.007)	-0.003 (0.008)	0.004 (0.008)	-0.003 (0.010)	-0.021*** (0.008)	-0.001 (0.011)	0.003 (0.008)	-0.003 (0.010)
Protestant	0.001 (0.012)		-0.008 (0.011)	-0.000 (0.011)	0.004 (0.010)	-0.001 (0.011)	-0.011 (0.011)	0.001 (0.012)	-0.002 (0.010)	0.018 (0.015)
Orthodox	0.018*** (0.006)	0.018*** (0.006)		0.018*** (0.006)	0.021*** (0.007)	0.018** (0.007)	0.011* (0.006)	0.021*** (0.006)	0.020*** (0.006)	0.028*** (0.009)
Other Christ	-0.004 (0.007)	-0.004 (0.007)	-0.004 (0.007)		-0.001 (0.007)	-0.003 (0.006)	0.001 (0.006)	0.003 (0.007)	-0.002 (0.007)	0.003 (0.007)
Muslim	-0.012 (0.008)	-0.012 (0.008)	-0.012 (0.008)	-0.012 (0.008)		-0.009 (0.009)	-0.017** (0.008)	-0.007 (0.008)	-0.010 (0.009)	-0.006 (0.009)
Jewish	0.000 (0.010)	0.000 (0.010)	0.000 (0.010)	0.000 (0.010)	0.002 (0.010)		-0.005 (0.010)	0.004 (0.010)	0.001 (0.010)	0.005 (0.011)
Buddhist	0.024*** (0.007)	0.024*** (0.007)	0.024*** (0.007)	0.029*** (0.008)	0.023*** (0.007)	0.021*** (0.007)		0.027*** (0.008)	0.024*** (0.007)	0.027*** (0.007)
Other Eastern	-0.007 (0.012)	-0.007 (0.012)	-0.002 (0.012)	-0.003 (0.011)	-0.007 (0.012)	-0.009 (0.011)	-0.007 (0.012)		-0.006 (0.011)	-0.003 (0.012)
Hindu	-0.010 (0.007)	-0.007 (-0.007)	-0.011 (0.008)	-0.007 (0.007)	-0.011 (0.007)	-0.012* (0.007)	-0.011 (0.007)	-0.008 (0.007)		-0.008 (0.007)
Other	-0.029* (0.017)	-0.030* (0.017)	-0.029* (0.017)	-0.026 (0.017)	-0.029* (0.017)	0.030* (0.017)	-0.029* (0.017)	-0.027 (0.017)	-0.029* (0.017)	
Constant	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)	2.668*** (0.272)

Table 3 (continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Observations	164	164	164	164	164	164	164	164	164	164
R^2	0.526	0.526	0.526	0.526	0.526	0.526	0.526	0.526	0.526	0.526
Adjusted R^2	0.488	0.488	0.488	0.488	0.488	0.488	0.488	0.488	0.488	0.488
Mean VIF	1.47	1.46	1.59	1.55	1.62	1.56	1.48	1.58	1.62	1.74
Cond. Number	33.48	33.48	33.48	33.48	33.48	33.48	33.48	33.48	33.48	33.48

Values in bold are coefficients for the log deviation from the geometric mean of all nine other religious components/categories. Dependent variable is log difference of per-capita GDP (PPP) between 2001 and 2020

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Closer inspection reveals that the bold diagonal coefficients in Table 3 are slightly larger than their counterparts in Table 2 by the same fixed factor. To see why, it is not necessary to write out the full elementwise form of $G_{d,h}^{-1}$. For the comparison of the highlighted diagonal entries in the two tables, it is enough to take $d = h$, so that the omitted ALR component and the isolated pivot-ILR component coincide.

The last row of $G_{h,h}^{-1}$, corresponding to the final pivot-ILR coefficient, is $\left[-\sqrt{\frac{D}{D-1}} 1'_{D-1} \ 0\right]$, where the final entry corresponds to the intercept. Hence $\bar{\beta}_{C/h}^{D-1} = -\sqrt{\frac{D}{D-1}} \sum_{i=1}^{D-1} \tilde{\beta}_{C/h}^i$. Using (15), this becomes $\bar{\beta}_{C/h}^{D-1} = \sqrt{\frac{D}{D-1}} \tilde{\beta}_{C/h}^h$. Therefore, the highlighted pivot-ILR coefficient for component h can be recovered from the corresponding ALR or log-contrast coefficient simply by multiplying by $\sqrt{D/(D-1)}$. By the same reasoning, the corresponding variance is multiplied by $D/(D-1)$, and the associated t -statistic is therefore unchanged.

Finally, collecting all D log elements within the $D-1$ pivot log-ratios in (21), the ILR component in (22) maps directly into the log-contrast coefficients in (9) where:

$$\check{\beta}_C^m = \begin{cases} -\sum_{r=1}^{D-1} \frac{\bar{\beta}_{C/1}^r}{\sqrt{r(r+1)}}, & m = 1, \\ \sqrt{\frac{m-1}{m}} \bar{\beta}_{C/1}^{m-1} - \sum_{r=m}^{D-1} \frac{\bar{\beta}_{C/1}^r}{\sqrt{r(r+1)}}, & m = 2, \dots, D-1, \\ \sqrt{\frac{D-1}{D}} \bar{\beta}_{C/1}^{D-1}, & m = D. \end{cases} \quad (23)$$

Some researchers advocate using additive log-ratios over isometric log-ratios on the grounds of simplicity of calculation and interpretation (Aitchison 2008; Greenacre et al. 2021). On the former point, inflating both the coefficients and standard errors associated with the composition variables by the factor $\sqrt{\frac{D}{D-1}}$ is sufficient to convert the regressions with ALR coordinates into those that would result from converting the data to D sets of pivot log-ratios and estimating each separately to obtain the coefficients in bold in Table 3. However, while these coefficients have a simple interpretation, the remaining coefficients in each column do not. As stated in Greenacre et al. (2023) on the use of ILRs “their major disadvantage is that their interpretation as single variables is complicated in all practical applications.”

6 Multicollinearity

Over recent decades, concern about multicollinearity as a problem in regression models has diminished. Nonetheless, to the degree that practitioners might be concerned, this section demonstrates that overall, replacing raw shares with log-ratios does not exacerbate the multicollinearity already implied by the intrinsic dependence among the shares in a composition.

Let $z_1, z_2, \dots, z_D$ be $n \times 1$ vectors of compositional variables. For each observation $\ell = 1, \dots, n$, the composition $(z_1^{(\ell)}, \dots, z_D^{(\ell)})$ follows a symmetric Dirichlet distribution, $(z_1^{(\ell)}, \dots, z_D^{(\ell)}) \sim \mathcal{D}(\alpha)$, with density $f(z_1^{(\ell)}, \dots, z_D^{(\ell)}) = \frac{\Gamma(D\alpha)}{\Gamma(\alpha)^D} \prod_{i=1}^D (z_i^{(\ell)})^{\alpha-1}$, where $z_i^{(\ell)} > 0$ for all i and $\sum_{i=1}^D z_i^{(\ell)} = 1$. With respect to Lebesgue measure, $E(z_i^{(\ell)}) = \frac{1}{D}$, $\text{Var}(z_i^{(\ell)}) = \frac{D-1}{D^2(1+D\alpha)}$, $\text{Cov}(z_i^{(\ell)}, z_j^{(\ell)}) = -\frac{1}{D^2(1+D\alpha)}$ for all $i \neq j$. Thus any two parts within a composition are negatively correlated: $\text{Corr}(z_i^{(\ell)}, z_j^{(\ell)}) = -\frac{1}{D-1}$, $i \neq j$, with the absolute value of this correlation decreasing in D .

For the corresponding $D-1$ log-ratios: $E(\log(z_i \oslash z_d)) = 0$, $\text{Var}(\log(z_i \oslash z_d)) = 2\varphi(\alpha)$ and $\text{Cov}(\log(z_i \oslash z_d), \log(z_j \oslash z_d)) = \varphi(\alpha)$, $i \neq j \neq d$, where $\varphi(\alpha) = \frac{\partial^2 \ln \Gamma(\alpha)}{\partial \alpha^2}$ is the trigamma function. The correlation between any pair of log-ratios is positive and fixed at $\text{Corr}(\log(z_i \oslash z_d), \log(z_j \oslash z_d)) = \frac{1}{2}$. Yet estimating with log-ratios, whether ALR or ILR, does not introduce greater multicollinearity as measured by the variance inflation factor (VIF).

Assume each vector z_i is length n and define the $n \times (D+1)$ matrix $Z = \{z_1, z_2, \dots, z_D, 1_n\}$ and the $n \times D$ matrix $Z_{\setminus d} = ZQ_{D,d}$, the matrix Z with the $d \in \{1, \dots, D\}$ column removed. Let $Z_{\setminus d \setminus i}$ denote the matrix obtained by removing both columns d and i from Z while retaining the intercept column, for $i \neq d$. Solving the expected normal equations:

$$Z'_{\setminus d \setminus i} Z_{\setminus d \setminus i} \hat{\beta}_{\setminus d \setminus i}^z = Z'_{\setminus d \setminus i} Z_i \quad (24)$$

where:

$$Z'_{\setminus d \setminus i} Z_{\setminus d \setminus i} = \frac{n}{D(1+D\alpha)} \begin{bmatrix} 1+\alpha & \alpha & \cdots & \alpha & 1+D\alpha \\ \alpha & 1+\alpha & & \vdots & \vdots \\ \vdots & & \ddots & \alpha & 1+D\alpha \\ \alpha & \cdots & \alpha & 1+\alpha & 1+D\alpha \\ 1+D\alpha & \cdots & \cdots & 1+D\alpha & D(1+D\alpha) \end{bmatrix},$$

$$Z'_{\setminus d \setminus i} Z_i = \frac{n}{D(1+D\alpha)} \begin{bmatrix} \alpha \\ \alpha \\ \vdots \\ \alpha \\ 1+D\alpha \end{bmatrix}$$

yields the vector of $D-1$ coefficients $(\hat{\beta}_{\setminus d \setminus i}^z)' = [-\frac{1}{2}, -\frac{1}{2}, \dots, -\frac{1}{2}, \frac{1}{2}]$, and the coefficient of determination of the regression of each vector z_i on the matrix $Z_{\setminus d \setminus i}$ is:

$$\hat{R}_{d,i}^2 = \frac{(\hat{\beta}_{\setminus d \setminus i}^z)' Z'_{\setminus d \setminus i} Z_{\setminus d \setminus i} \hat{\beta}_{\setminus d \setminus i}^z - nE(z_i)^2}{z_i' z_i - nE(z_i)^2} = \frac{D-2}{2(D-1)}. \quad (25)$$

For aregression with $Z_{\setminus d}$ as the only explanatory variables, $VIF = 1 / (1 - \hat{R}_{d,i}^2) = 2 \frac{D-1}{D}$.

Similarly the matrix $\tilde{Z}_{/d \setminus i}$ is the matrix $\tilde{Z}_{/d}$ of additive log-ratios with respect to z_d , and an additional column of ones with the $i \neq d$ column removed. Solving the normal equations:

$$\tilde{Z}'_{/d \setminus i} \tilde{Z}_{/d \setminus i} \tilde{\beta}_{/d \setminus i}^z = \tilde{Z}'_{/d \setminus i} \tilde{z}_i \quad (26)$$

where:

$$\tilde{Z}'_{/d \setminus i} \tilde{Z}_{/d \setminus i} = n \begin{bmatrix} 2\varphi(\alpha) & \varphi(\alpha) & \cdots & \varphi(\alpha) & 0 \\ \varphi(\alpha) & 2\varphi(\alpha) & \cdots & \varphi(\alpha) & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \varphi(\alpha) & \varphi(\alpha) & \cdots & 2\varphi(\alpha) & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix},$$

$$\tilde{Z}'_{/d \setminus i} \tilde{z}_i = n \begin{bmatrix} \varphi(\alpha) \\ \varphi(\alpha) \\ \vdots \\ \varphi(\alpha) \\ 0 \end{bmatrix}$$

yields the vector of $D - 1$ coefficients $(\tilde{\beta}_{/d \setminus i}^z)' = [1/(D - 1), 1/(D - 1), \dots, 1/(D - 1), 0]$. Though the matrices appear very different, applying the same formula as in (25) yields the same coefficient of determination $\tilde{R}_{d,i}^2$, and hence the same VIF.

Nonetheless, the projection matrices associated with removing an additional column from either $\tilde{C}_{/d}$ or $\tilde{C}_{/h}$ are not permutation-invariant. Hence the auxiliary regressions used to compute variance inflation factors are not permutation-invariant either, and so the mean VIFs differ across the alternative ALR representations in Table 2. Since the different sets of regressions using ALR or ILR coordinates each map to the same underlying log-contrast models, these differences provide no basis for preferring one permutation over another.

By contrast, the condition numbers computed from the centered, unscaled matrices of regressors corresponding to the ALR regressions in Table 2 are not permutation-invariant because the matrices linking alternative ALR representations are not orthogonal. The salient component is H_j in (12), whose j th row, containing the vector $-1'_{D-1}$, is neither orthogonal to the remaining rows nor of unit length. Hence H_j does not preserve singular values. In the ILR case, however, the transformations linking the alternative orthonormal bases in Table 3 are generated by products of orthogonal permutation matrices S_j . Hence singular values, and therefore the condition numbers, are invariant across the different ILR representations. Even so, that property alone is insufficient reason to prefer the more complicated ILR coordinates to ALRs in regression analysis.

7 Conclusion

ALR regressions are simple to estimate and interpret, whereas ILR regressions appear in some work because their coordinates are defined with respect to an orthogonal basis. In both cases, the data enter the regression as coordinates in a vector space.

Logarithmic ratios are now commonly used in geophysics and biology. A small number of papers in the economics literature use log-ratios in regression analysis, including work by Fry et al. (1996); Kynčlová et al. (2015); Creixans-Tenas et al. (2019). Although the examples in this paper focus on growth regressions, the approach could be used more widely in the empirical social sciences.

Appendix A. Proofs

Appendix A.1. Proof of Lemma 1

$$\begin{aligned}
 P_{C \setminus d} &= C_{\setminus d} \left(C'_{\setminus d} C_{\setminus d} \right)^{-1} C'_{\setminus d} = C_{\setminus f} L_{d,f} \left((C_{\setminus f} L_{d,f})' C_{\setminus f} L_{d,f} \right)^{-1} (C_{\setminus f} L_{d,f})' \\
 &= C_{\setminus f} L_{d,f} L_{d,f}'^{-1} \left(C'_{\setminus f} C_{\setminus f} \right)^{-1} (L'_{d,f})^{-1} L'_{d,f} C'_{\setminus f} \\
 &= C_{\setminus f} \left(C'_{\setminus f} C_{\setminus f} \right)^{-1} C'_{\setminus f} = P_{C \setminus f} = P_C \\
 P_{X \setminus d} &= X_{\setminus d} \left(X'_{\setminus d} X_{\setminus d} \right)^{-1} X'_{\setminus d} \\
 &= P_N + (I_n - P_N) C_{\setminus d} \left\{ [(I_n - P_N) C_{\setminus d}]' (I_n - P_N) C_{\setminus d} \right\}^{-1} [(I_n - P_N) C_{\setminus d}]' \\
 &= P_N + (I_n - P_N) C_{\setminus f} L_{d,f} \left\{ [(I_n - P_N) C_{\setminus f} L_{d,f}]' (I_n - P_N) C_{\setminus f} L_{d,f} \right\}^{-1} \\
 &\quad L'_{d,f} C'_{\setminus f} (I_n - P_N)' \\
 &= P_N + (I_n - P_N) C_{\setminus f} L_{d,f} (L_{d,f})^{-1} \left\{ [(I_n - P_N) C_{\setminus f}]' (I_n - P_N) C_{\setminus f} \right\}^{-1} (L'_{d,f})^{-1} \\
 &\quad L'_{d,f} C'_{\setminus f} (I_n - P_N)' \\
 &= P_N + (I_n - P_N) C_{\setminus f} \left\{ [(I_n - P_N) C_{\setminus f}]' (I_n - P_N) C_{\setminus f} \right\}^{-1} [(I_n - P_N) C_{\setminus f}]' \\
 &= X_{\setminus f} \left(X'_{\setminus f} X_{\setminus f} \right)^{-1} X'_{\setminus f} = P_{X \setminus f} = P_X
 \end{aligned}$$

Replacing $P_{X \setminus d}$ with P_X in (6): $\hat{\varepsilon}_{\setminus d} = \hat{\varepsilon}_{\setminus f} = \hat{\varepsilon}$.

Appendix A.2 Proof of Theorem 2

Define: $W_{\setminus d} \equiv C_{\setminus d} Q_{D-1,D} = C Q_{D,d} Q_{D-1,D}$ and $M \equiv [N; 1_n]$,
 $P_M \equiv M(M'M)^{-1}M'$ and $X_{\setminus d} = [M; W_{\setminus d}]$.

$$\begin{aligned}
 F[D-1, n-K-D] &= \frac{\mathbf{y}'(\mathbf{I}_n - \mathbf{P}_M)\mathbf{W}_{\setminus d} \left(\mathbf{W}_{\setminus d}'(\mathbf{I}_n - \mathbf{P}_M)\mathbf{W}_{\setminus d} \right)^{-1} \mathbf{W}_{\setminus d}'(\mathbf{I}_n - \mathbf{P}_M)\mathbf{y}/(D-1)}{\mathbf{y}'(\mathbf{I}_n - \mathbf{P}_X)\mathbf{y}/(n-K-D)} \\
 &= \frac{\mathbf{y}'(\mathbf{P}_X - \mathbf{P}_M)\mathbf{y}/(D-1)}{\mathbf{y}'(\mathbf{I}_n - \mathbf{P}_X)\mathbf{y}/(n-K-D)}.
 \end{aligned}$$

Where $\mathbf{P}_X$ and $\mathbf{P}_M$ are permutation invariant

Appendix A.3 Proof of Theorem 3

The permutation invariance of $\hat{\beta}_N$ and $\hat{\epsilon}$ together with equation (3) implies $\mathbf{C}_{\setminus d}\hat{\beta}_{C\setminus d} = \mathbf{C}_{\setminus f}\hat{\beta}_{C\setminus f}$. Premultiplying both sides by $\mathbf{C}_{\setminus d}'$ and solving for $\hat{\beta}_{C\setminus d}$:

$$\hat{\beta}_{C\setminus d} = \left(\mathbf{C}_{\setminus d}'\mathbf{C}_{\setminus d} \right)^{-1} \mathbf{C}_{\setminus d}'\mathbf{C}_{\setminus f}\hat{\beta}_{C\setminus f} = \left(\mathbf{C}_{\setminus d}'\mathbf{C}_{\setminus d} \right)^{-1} \mathbf{C}_{\setminus d}'\mathbf{C}_{\setminus d}\mathbf{L}_{f,d}\hat{\beta}_{C\setminus f} = \mathbf{L}_{f,d}\hat{\beta}_{C\setminus f}$$

$$\mathbf{L}_{f,d} \text{ is non-random: } \text{Var} \left(\hat{\beta}_{C\setminus d} \right) = \text{Var} \left(\mathbf{L}_{f,d}\hat{\beta}_{C\setminus f} \right) = \mathbf{L}_{f,d}\text{Var} \left(\hat{\beta}_{C\setminus f} \right) \mathbf{L}_{f,d}'$$

Appendix A.4 Proof of Lemma 2

$$\begin{aligned}
 \tilde{\mathbf{P}}_{C/d} &= \tilde{\mathbf{C}}_{/d} \left(\tilde{\mathbf{C}}_{/d}'\tilde{\mathbf{C}}_{/d} \right)^{-1} \tilde{\mathbf{C}}_{/d}' = \tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f} \left(\left(\tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f} \right)' \tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f} \right)^{-1} \left(\tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f} \right)' \\
 &= \tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f}\mathbf{M}_{d,f}'^{-1} \left(\tilde{\mathbf{C}}_{/f}'\tilde{\mathbf{C}}_{/f} \right)^{-1} \left(\mathbf{M}_{d,f}' \right)^{-1} \mathbf{M}_{d,f}'\tilde{\mathbf{C}}_{/f}' \\
 &= \tilde{\mathbf{C}}_{/f} \left(\tilde{\mathbf{C}}_{/f}'\tilde{\mathbf{C}}_{/f} \right)^{-1} \tilde{\mathbf{C}}_{/f}' = \tilde{\mathbf{P}}_{C/f} = \tilde{\mathbf{P}}_C \\
 \tilde{\mathbf{P}}_{X/d} &= \tilde{\mathbf{X}}_{/d} \left(\tilde{\mathbf{X}}_{/d}'\tilde{\mathbf{X}}_{/d} \right)^{-1} \tilde{\mathbf{X}}_{/d}' \\
 &= \mathbf{P}_N + (\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/d} \left\{ \left[(\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/d} \right]' (\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/d} \right\}^{-1} \left[(\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/d} \right]' \\
 &= \mathbf{P}_N + (\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f} \left\{ \left[(\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f} \right]' (\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f} \right\}^{-1} \\
 &\quad \mathbf{M}_{d,f}'\tilde{\mathbf{C}}_{/f}'(\mathbf{I}_n - \mathbf{P}_N)' \\
 &= \mathbf{P}_N + (\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f}\mathbf{M}_{d,f}(\mathbf{M}_{d,f}')^{-1} \\
 &\quad \left\{ \left[(\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f} \right]' (\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f} \right\}^{-1} \\
 &\quad (\mathbf{M}_{d,f}')^{-1}\mathbf{M}_{d,f}'\tilde{\mathbf{C}}_{/f}'(\mathbf{I}_n - \mathbf{P}_N)' \\
 &= \mathbf{P}_N + (\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f} \left\{ \left[(\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f} \right]' (\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f} \right\}^{-1} \left[(\mathbf{I}_n - \mathbf{P}_N)\tilde{\mathbf{C}}_{/f} \right]' \\
 &= \tilde{\mathbf{X}}_{/f} \left(\tilde{\mathbf{X}}_{/f}'\tilde{\mathbf{X}}_{/f} \right)^{-1} \tilde{\mathbf{X}}_{/f}' = \tilde{\mathbf{P}}_{X/f} = \tilde{\mathbf{P}}_X
 \end{aligned}$$

The error terms are: $\tilde{\epsilon} = (\mathbf{I}_n - \tilde{\mathbf{P}}_X)\mathbf{y}$.

Appendix A.5 Proof of Lemma 3

$$\begin{aligned}
 \bar{P}_{C/h} &= \bar{C}_{/h} \left(\bar{C}'_{/h} \bar{C}_{/h} \right)^{-1} \bar{C}'_{/h} = \tilde{C}_{/d} G_{d,h} \left(G'_{d,h} \tilde{C}'_{/d} \tilde{C}_{/d} G_{d,h} \right)^{-1} G'_{d,h} \tilde{C}'_{/d} \\
 &= \tilde{C}_{/d} G_{d,h} G_{d,h}^{-1} \left(\tilde{C}'_{/d} \tilde{C}_{/d} \right)^{-1} \left(G'_{d,h} \right)^{-1} G'_{d,h} \tilde{C}'_{/d} = \tilde{C}_{/d} \left(\tilde{C}'_{/d} \tilde{C}_{/d} \right)^{-1} \tilde{C}'_{/d} = \tilde{P}_C \\
 \bar{P}_{X/h} &= \bar{X}_{/h} \left(\bar{X}'_{/h} \bar{X}_{/h} \right)^{-1} \bar{X}'_{/h} \\
 &= P_N + (I_n - P_N) \bar{C}_{/h} \left\{ [(I_n - P_N) \bar{C}_{/h}]' (I_n - P_N) \bar{C}_{/h} \right\}^{-1} [(I_n - P_N) \bar{C}_{/h}]' \\
 &= P_N + (I_n - P_N) \tilde{C}_{/d} G_{d,h} \left\{ [(I_n - P_N) \tilde{C}_{/d} G_{d,h}]' (I_n - P_N) \tilde{C}_{/d} G_{d,h} \right\}^{-1} \\
 &\quad G'_{d,h} \tilde{C}'_{/d} (I_n - P_N)' \\
 &= P_N + (I_n - P_N) \tilde{C}_{/d} G_{d,h} (G_{d,h})^{-1} \left\{ [(I_n - P_N) \tilde{C}_{/d}]' (I_n - P_N) \tilde{C}_{/d} \right\}^{-1} \\
 &\quad (G'_{d,h})^{-1} G'_{d,h} \tilde{C}'_{/d} (I_n - P_N)' \\
 &= P_N + (I_n - P_N) \tilde{C}_{/d} \left\{ [(I_n - P_N) \tilde{C}_{/d}]' (I_n - P_N) \tilde{C}_{/d} \right\}^{-1} [(I_n - P_N) \tilde{C}_{/d}]' \\
 &= \tilde{X}_{/d} \left(\tilde{X}'_{/d} \tilde{X}_{/d} \right)^{-1} \tilde{X}'_{/d} = \tilde{P}_{X/d} = \tilde{P}_X
 \end{aligned}$$

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00180-026-01774-y>.

Funding Not applicable.

Data availability The data used in the examples can be downloaded from <https://osf.io/btf4h/>.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

Ethical approval This article does not contain any studies with human participants or animals performed by the author.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitchison J (1982) The statistical analysis of compositional data. *J Roy Stat Soc: Ser B (Methodol)* 44:139–160. <https://doi.org/10.1111/j.2517-6161.1982.tb01195.x>
- Aitchison J (1986) The statistical analysis of compositional data. Monographs on Statistics and Applied Probability. Chapman and Hall London. <https://doi.org/10.1007/978-94-009-4109-0> (Reprinted in 2003)
- Aitchison J (2008) The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. <https://www.semanticscholar.org/paper/The-single-principle-of-compositional-data-and-some-Aitchison/f106efc3e9b4889af2482a788b5886e53789c301>. keynote address, CODAWORK08
- Aitchison J, Bacon-Shone J (1984) Log contrast models for experiments with mixtures. *Biometrika* 71:323–330. <https://doi.org/10.1093/biomet/71.2.323>
- Anagol S, Fujiwara T (2016) The runner-up effect. *J Polit Econ* 124:927–991. <https://doi.org/10.1086/686746>
- Barro RJ, McCleary RM (2003) Religion and economic growth. *Am Sociol Rev* 68:760–781. <https://doi.org/10.2307/1519761>
- Berman N, Couttenier M, Rohner D, Thoenig M (2017) This mine is mine! how minerals fuel conflicts in Africa. *Am Econ Rev* 107:1564–1610. <https://doi.org/10.1257/aer.20150774>
- Billheimer D, Guttorp P, Fagan WF (2001) Statistical interpretation of species composition. *J Am Stat Assoc* 96:1205–1214. <https://doi.org/10.1198/016214501753381850>
- Brodeur A, Lé M, Sangnier M, Zylberberg Y (2016) Star wars: the empirics strike back. *Am Econ J Appl Econ* 8:1–32. <https://doi.org/10.1257/app.20150044>
- Burke MA, Sass TR (2013) Classroom peer effects and student achievement. *J Law Econ* 31:51–82. <https://doi.org/10.1086/666653>
- Catherine S, Sodini P, Zhang Y (2024) Countercyclical income risk and portfolio choices: evidence from Sweden. *J Financ* 79:1755–1788. <https://doi.org/10.1111/jofi.13341>
- Chen J, Zhang X, Li S (2017) Multiple linear regression with compositional response and covariates. *J Appl Stat* 44:2270–2285. <https://doi.org/10.1080/02664763.2016.1157145>
- Coenders G, Pawlowsky-Glahn V (2020) On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT-Stat Oper Res Trans*: 201–220. <https://doi.org/10.2436/20.80.80.02.100>
- Creixans-Tenas J, Coenders G, Arimany-Serrat N (2019) Corporate social responsibility and financial profile of spanish private hospitals. *Heliyon* 5. <https://doi.org/10.1016/j.heliyon.2019.e02623>
- Devarajan S, Swaroop V, Zou, HF (1996) The composition of public expenditure and economic growth. *J Monet Econ* 37:313–344. [https://doi.org/10.1016/S0304-3932\(96\)90039-2](https://doi.org/10.1016/S0304-3932(96)90039-2)
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35:279–300. <https://doi.org/10.1023/A:1023818214614>
- Fry JM, Fry TRL, McLaren KR (1996) The stochastic specification of demand share equations: restricting budget shares to the unit simplex. *J Econ* 73:377–385. [https://doi.org/10.1016/S0304-4076\(95\)01727-5](https://doi.org/10.1016/S0304-4076(95)01727-5)
- Greenacre M (2018) Compositional data analysis in practice. CRC Press. <https://doi.org/10.1201/9780429455537>
- Greenacre M, Grunsky E, Bacon-Shone J (2021) A comparison of isometric and amalgamation logratio balances in compositional data analysis. *Comput Geosci* 148. <https://doi.org/10.1016/j.cageo.2020.104621>
- Greenacre M, Grunsky E, Bacon-Shone J, Erb I, Quinn T (2023) Aitchison's compositional data analysis 40 years on: a reappraisal. *Stat Sci* 38:386–410
- Hall RE, Jones CI (1999) Why do some countries produce so much more output per worker than others? *Q J Econ* 114:83–116. <https://doi.org/10.1162/003355399555954>
- Hamm L, McDonald S (2015) Helping hands: race, neighborhood context, and reluctance in providing job-finding assistance. *Sociol Q* 56:539–557. <https://doi.org/10.1111/tsq.12091>
- Hron K, Filzmoser P, Thompson K (2012) Linear regression with compositional explanatory variables. *J Appl Stat* 39:1115–1128. <https://doi.org/10.1080/02664763.2011.644268>
- Kynčlová P, Filzmoser P, Hron K (2015) Modeling compositional time series with vector autoregressive models. *J Forecast* 34:303–314. <https://doi.org/10.1002/for.2336>
- Lindh T (1999) Age structure and economic policy: the case of saving and growth. *Popul Res Policy Rev* 18:261–277. <https://doi.org/10.1023/A:1006123514049>

- Logan JR, Minca E, Adar S (2012) The geography of inequality: why separate means unequal in American public schools. *Sociol Educ* 85:287–301. <https://doi.org/10.1177/0038040711431588>
- Mankiw NG, Romer D, Weil DN (1992) A contribution to the empirics of economic growth. *Q J Econ* 107:407–437. <https://doi.org/10.2307/2118477>
- Maoz Z, Henderson EA (2013) The world religion dataset, 1945–2010: Logic, estimates, and trends. *Int Interact* 39:265–291. <https://doi.org/10.1080/03050629.2013.782306>
- McCleary RM, Barro RJ (2006) Religion and economy. *J Econ Perspect* 20:49–72. <https://doi.org/10.1257/jep.20.2.49>
- Noland M (2005) Religion and economic performance. *World Dev* 33:1215–1232. <https://doi.org/10.1016/j.worlddev.2005.03.006>
- Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stoch Env Res Risk Assess* 15:384–398. <https://doi.org/10.1007/s004770100077>
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley. <https://doi.org/10.1002/9781119003144>
- Rodrik D, Subramanian A, Trebbi F (2004) Institutions rule: the primacy of institutions over geography and integration in economic development. *J Econ Growth* 9:131–165. <https://doi.org/10.1023/B:JOEG.0000031425.72248.85>
- Trounstine J (2016) Segregation and inequality in public goods. *Am J Polit Sci* 60:709–725. <https://doi.org/10.1111/ajps.12227>
- Voigt S, Gutmann J, Feld LP (2015) Economic growth and judicial independence, a dozen years on: cross-country evidence using an updated set of indicators. *Eur J Polit Econ* 38:197–211. <https://doi.org/10.1016/j.ejpolco.2015.01.004>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.