HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

THE UNIVERSITY of York

WP 23/16

Use of compositional covariates in linear regression: problems and solutions

Tianchang Zhao; Matt Sutton and Rachel Meacock

November 2023

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

Use of compositional covariates in linear regression: problems and solutions

Tianchang Zhao[*][†]
Matt Sutton[†]
Rachel Meacock[†]

Abstract

Compositional variables such as proportions by age group are commonly included as covariates in aggregate-level health research. Since these proportions sum to one and only contain relative information, directly including them as covariates violates the fundamental assumptions made in linear regression analysis. We explain the compositional nature of such data and, using practice-level elective admissions rates in England as an example outcome variable, demonstrate the consequences of directly using proportions in regressions. We also provide an overview of compositional data analysis (CoDA) techniques with a focus on isometric log-ratio (ILR) transformation. Applying ILR to our example data shows that the regression results can differ significantly from those obtained using raw proportions. Health economists should apply appropriate CoDA methods when using compositional data in their research.

Keywords: compositional data, CoDA, age group proportion, isometric log-ratio
JEL: C18, C13, I10

[*] Corresponding author: tianchang.zhao@postgrad.manchester.ac.uk
[†] Health Organisation, Policy and Economics, University of Manchester

Introduction

Compositional variables are variables that sum to a constant. For example, the proportion of patients in different age bands (which sum to 1), or the time spent on different activities during a day (which sum to 24 hours). The components of compositional data are mutually exclusive and exhaustive. Compositional variables are singular as there exists a perfect linear relationship between the parts. As a result, compositional variables only contain relative information contained in the ratios between the parts, instead of absolute information. The sum of the compositions can therefore be chosen arbitrarily. For example, proportions ($\sum_{i=1}^{D} x_i = 1$, where x are the components or parts and D is the total number of parts) can be changed into percentage points ($\sum_{i=1}^{D} x_i = 100$) in regression analysis. Another key property of compositional variables is co-dependence: it is impossible to change one of the components without changing at least one other because of the constant sum restriction.

Since compositional variables are by definition co-dependent and finite, applying any statistical methods that use the covariance or correlation matrix of the variables to such compositional data without considering the special geometrical properties is inappropriate (Pawlowsky-Glahn and Egozcue, 2006). Mathematically speaking, standard statistical techniques are developed for data in the real space, following Euclidean geometry, while compositional data exist in the simplex space following the Aitchison geometry (Aitchison, 1982). Compositional variables $\mathbf{x}$ of D mutually exclusive and exhaustive parts that sum to C have the following simplex sample space S^D:

$$S^D = \left\{ \mathbf{x} = (x_1, \dots, x_d)', x_i > 0, \sum_{i=1}^{D} x_i = C \right\}.$$

Conventional techniques including linear regression analysis which require multivariate normality and limited multicollinearity of variables should not be applied to the raw form of compositional data because compositional data violate the basic assumptions made by these techniques, e.g., the variables should be free to vary in the real space (Pawlowsky-Glahn and Egozcue, 2006). Such practice can produce misleading results, showing spurious effects or obscuring genuine effects, and the estimated coefficients of non-compositional covariates in the model are also affected, producing biased estimates (Pearson, 1897; Vistelius and Sarmanov, 1961; Filzmoser et al., 2018).

The interpretation of the estimated coefficients on compositional variables included in their raw form is also problematic. The standard interpretation of regression coefficients is the effect of a one-unit increase of the variable of interest while holding the values of all other variables constant, which is impossible for compositional variables. The issues outlined above apply to the inclusion of compositional variables in regression models even when these are only included as “control” variables, since the estimated coefficients of other non-compositional covariates are nevertheless impacted, regardless of whether the coefficients on the compositional variables are to be interpreted.

Problems with compositional data analysis were first highlighted formally in the 1960s (Chayes, 1960; Chayes, 1962), and Rock (1988) provides a comprehensive description of six problems of applying conventional statistic methods to compositional data. These issues were mostly neglected due to the absence of feasible solutions, until

1982 when Aitchison created Aitchison geometry for the simplex space and proposed log-contrast methods, additive log-ratio (alr) and centred log-ratio (clr) transformations, to allow meaningful regression analysis of compositions by transforming them into coordinates in the standard real space (Aitchison, 1982). Most of the progress in the statistical analysis of compositional data achieved in the last three decades are based on Aitchison's contribution. Log-ratio transformations are now applied as standard practice in studies of geology (Buccianti et al., 2006), chemistry (Arnold et al., 2020), microbiome (Gloor et al., 2017), and epidemiology (Mert et al., 2018), where the compositional variables are often the main variables of interest. However, such techniques are not common practice in health economics research.

In this paper we aim to demonstrate the properties of compositional data in the context of health economics, using the elective hospital admission rate of the population at the GP practice level as an example outcome variable and proportions of GP practice's registered patients in different age groups as explanatory variables. The proportions of patients in different age groups are often included as covariates in aggregate level health research to describe the age structure of the unit of analysis, such as a geographical region or a healthcare organization. Since the age proportions have a fixed and constant sum constraint of 1 or 100 percentage points for all units, i.e., $\sum_{i=1}^{d} prop_i = 1$, where d is the number of age bands, they are compositional data. Many other variables of compositional nature may be of interest to health economists, such as the proportion of budget spent on healthcare, and time spent on different activities during a day (e.g. Drastichová and Filzmoser, 2020; Urwin et al., 2023).

This paper aims to raise awareness of the issues of using variables of compositional nature in linear regression models, as is typically used in heath economics. We first demonstrate the consequences of directly using percentages as explanatory variables in linear regression models, but the issues also exist for other models developed for real data, linear or nonlinear. We then introduce basic compositional data analysis (CoDA) methods and demonstrate the procedures of applying isometric log-ratio (ilr) transformation to the age proportions data. Finally, we compare the regression results obtained with and without the transformation.

Data

We obtained the dataset used in Gibson et al. (2022) from its authors. It contains 2019 annual data on elective hospital admissions amongst the populations of patients registered at 6,212 GP practices in England, taken from the Hospital Episode Statistics (HES) database, and registered patient and workforce data from the NHS Digital GP Workforce Dataset (NHS Digital, 2023b). The dataset contains the numbers of registered patients at each general practice in seven different age groups. For clarity and convenience of visual demonstration in our example, we merged the population counts into three age groups: 0-14, 15-64, and 65+, but the issues we demonstrate apply to all compositional data regardless of the number of parts. The proportion of patients in each age group is calculated as the number of patients in that age group divided by the total number of registered patients of the practice, and these are expressed as percentage points. The sample space of the age proportion data is the simplex:

$$\mathcal{S}^3 = \left\{ P = (prop_{\text{0-14}}, prop_{\text{15-64}}, prop_{65+}), prop_i > 0, \sum_{i=1}^{3} prop_i = 100\% \right\}.$$

The number of annual elective hospital admissions per registered patient is used as the outcome variable. As well as the compositional age variables, we include the following variables as covariates in the model:

- Deprivation: income deprivation of the area in which a practices' registered population resides, weighted according to the Lower-layer Super Output Areas (LSOA) of each registered patient. The deprivation scores for each LSOA are taken from the English indices of deprivation 2019.
- Rural location: a binary indicator of practice rurality from the NHS Payments to Practices datasets (2020).
- Access: practice-level measure of average time since last seen a GP, calculated using GP Patient Survey results and methods described in Gibson et al. (2022).

Descriptive statistics

Due to the special geometric properties discussed in the Introduction section, standard measures that rely on Euclidean geometry, including arithmetic mean, variance, standard deviation, and covariance, should not be applied to compositional data. The most appropriate measures of central tendency, spread, and codependence will differ depending upon the characteristics of the data, but meaningful measures of compositional data should satisfy scaling invariance, perturbation invariance, permutation invariance, and subcompositional coherence (Aitchison, 1982; Van den Boogaart and Tolosana-Delgado, 2013).

To measure the central tendency of our example data, a composition that best represents the centre of the proportions, the compositional geometric mean, can be obtained by first calculating the geometric mean of the proportions, then dividing by the sum of the three geometric means (Chastin and Palarea-Albaladejo, 2015). For example, the geometric mean of the proportion of patients aged 0-14 among all practices is:

$$g_{0\text{-}14} = \sqrt[6212]{\prod_{i=1}^{6212} prop_{0\text{-}14,\,i}}$$

where i is the practice. This can also be expressed as:

$$g_{0\text{-}14} = e^{\frac{\sum_{i=1}^{6212} \ln prop_{0\text{-}14,i}}{6212}}$$

i.e., the logarithm of the geometric mean is simply the arithmetic mean of the logged numbers. Then, the compositional geometric mean of the youngest age group, age 0-14, is:

$$\overline{prop_{0\text{-}14}} = \frac{g_{0\text{-}14}}{g_{0\text{-}14} + g_{15\text{-}64} + g_{65+}} = \frac{16.76}{16.76 + 64.77 + 15.94} = 0.1719$$

The compositional geometric mean vector of our dataset is $(0.1719, 0.6645, 0.1635)$, representing the "centre" of the distribution of age proportions of the practices. This can be obtained using the `mean.acomp(x)` function in R package '`compositions`' (van den Boogaart et al., 2013).

The dispersion of compositions can be measured by its metric variance, or total variance (Pawlowsky-Glahn and Egozcue, 2001), defined as the (degree of freedom corrected) average squared distance of the composition to its centre. In our example, the metric variance of the proportions is calculated using the `mvar(x)` function in R package '`compositions`' as

$$\text{mvar}(prop_{0\text{-}14}, prop_{15\text{-}64}, prop_{65+}) = \frac{1}{3-1}\sum_{i=1}^{3} d^2(prop_i, \overline{prop}) = 0.2771$$

The covariance matrix of the proportions cannot be used as a measure of co-dependence because the covariances are spurious as a result of the constant sum constraint. However, a variation matrix can be constructed by calculating the variances of log-ratios between each pair of proportions, i.e., $Var(\ln\frac{x_i}{x_j})$ (Aitchison, 1982). The variation matrix can be estimated using the `variation(x)` function in R package '`compositions`' and is shown in Table 1. Each element of the matrix τ_{kj} is estimated by:

$$\hat{\tau}_{kj} = \frac{1}{6212-1}\sum_{i=1}^{6212} \ln^2\frac{x_{ik}}{x_{ij}} - \ln^2\frac{\bar{x}_k}{\bar{x}_j}$$

The closer this number is to zero, the higher the level of co-dependence between the two variables. Table 1 shows that the co-dependence between proportions of 0-14 and 15-64 is much stronger than that between children and older people (0-14 and 65+).

Table 1 Variation matrix of proportions of patients in three age groups

	Prop 0to14	Prop 15to64	Prop 65+
Prop 0to14	0	0.1005	0.3545
Prop 15to64	0.1005	0	0.3762
Prop 65+	0.3545	0.3762	0

Compositional graphics

Because the compositional variables are co-dependent, conventional visual representations of data can be highly misleading. For example, Figure 1 plots the elective hospital admission rates and the percentage of patients in each of the three age groups. The figure suggests that practices with higher proportions of patients aged 65 years and over tend to have higher admission rates, and practices with higher proportions of patients aged 15-64 years tend to have lower elective admission rates. Whilst these results appear sensible, this way of presenting the proportions ignores the fact that any changes in one of the plots would necessarily be accompanied by changes of the same total amount in the other plots. This is more obvious when we plot the proportions against each other, as shown in Figure 2.

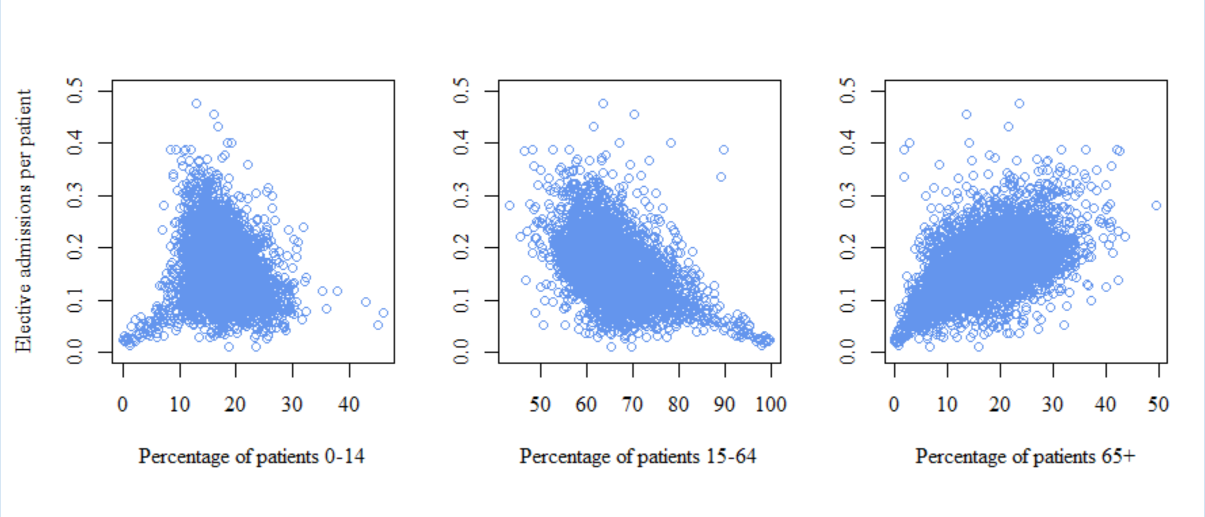

Figure 1 Elective admission rate vs proportions of patients in three age groups

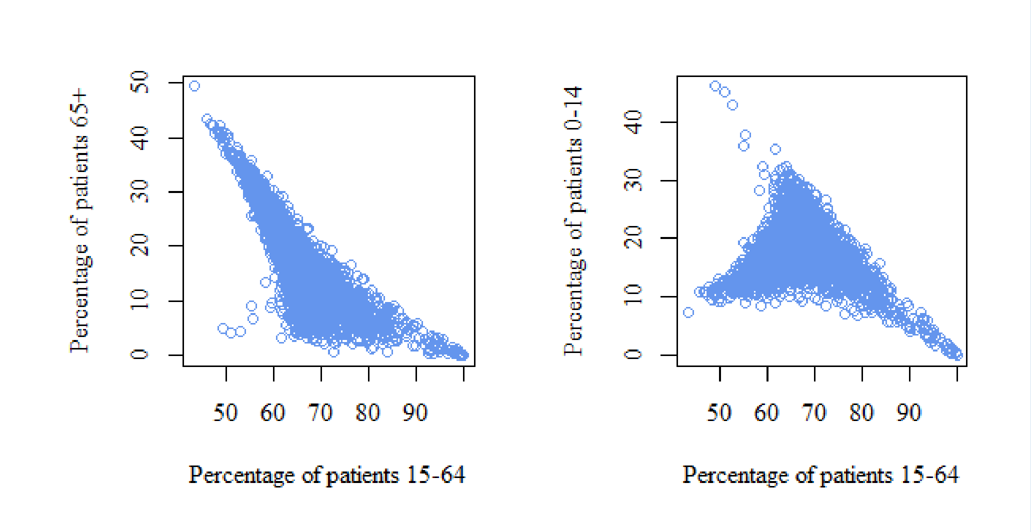

Figure 2 Percentage of patients 65+ and 0-14 against percentage of patients 15-64

Since the sample space of the proportions of patients in the three age groups is the S^3 simplex space, depicted by the shaded two-dimensional triangle plane in a three-dimensional space as shown in Figure 3, it is possible to plot the proportions on this triangle with each side annotated by one of the axes. This results in a ternary diagram (Figure 4), which emphasizes the fact that age proportions are co-dependent, and one proportion cannot change without also changing at least one of the other proportions. The plot on the left of Figure 4 is on the full 0 to 100 percentage points scale, and the plot on the right zooms in to show more details. The ternary diagram is arguably the most intuitive and straightforward way to show the compositions of three parts in their raw form. To visually represent compositions of more than three parts, compositional biplots, scatterplots of log-ratios, and sequences of bar plots are sensible options, although they are not as intuitive as the ternary plots (Aitchison, 2008; Van den Boogaart and Tolosana-Delgado, 2013).

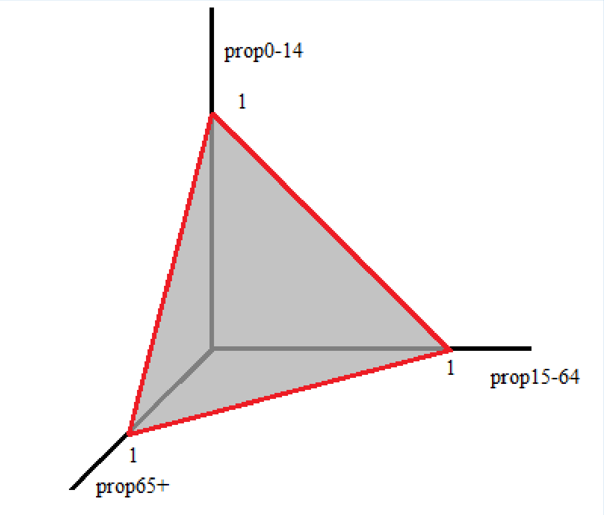

Figure 3 sample space of the three proportions

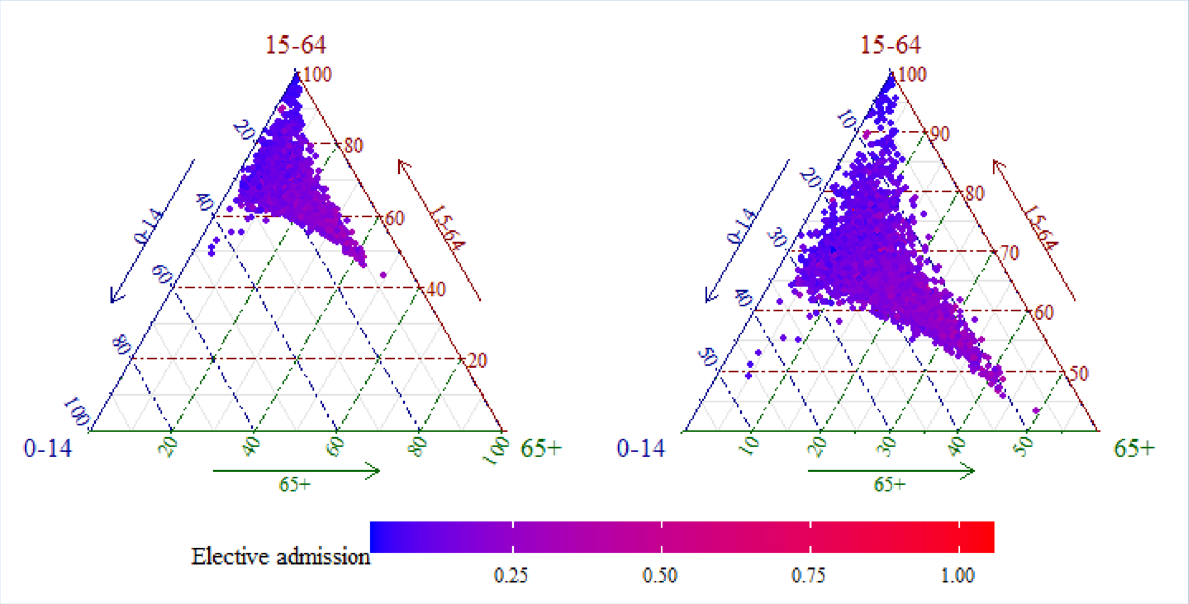

Figure 4 Ternary plot of practice age group proportions on different scales

Linear regression with age proportions

The plots in Figure 1 appear to show linear relationships between proportions (expressed as percentage points) of certain age groups and the elective hospital admission rate. However, directly using these age proportions as explanatory variables in a linear regression model may produce misleading results. To show this, we regress the elective hospital admission rate against the three proportions. Since the proportions sum to 100%, including all three in the model causes perfect singularity and statistical software such as R and Stata would arbitrarily omit one of the variables. Due to the constant sum constraint, any pair of two proportions contain the exact same information as contained by all three proportions. Running three versions of the following model:

$$\text{Elective admission rate} = \beta_0 + \beta_1 prop_{0-14} + \beta_2 prop_{15-64} + \beta_3 prop_{65+} + e$$

by omitting one of the proportions each time, the R^2 of each model would be exactly the same, as shown in Table 2. Practices with lower than 1% of registered patients in any of the three age groups are considered outliers and removed from the dataset as they may suffer from data quality issue, leaving 6,190 GP practices in our main analysis sample.

Three regression results are shown in Table 2, each omitting one of the proportions. Notice that the coefficient of $prop_{65+}$ in model (1) is the same as that of $prop_{15-64}$ in model (2), the coefficient of $prop_{0-14}$ in (2) is the same as $prop_{65+}$ in (3), and the coefficient of $prop_{15-64}$ in model (3) is the same as that of $prop_{0-14}$ in model (1), only with opposite signs in each case. The coefficients within each of these pairs also display identical standard errors.

Table 2 Regression of the elective admission rate on each pair of age group proportions

	(1)	(2)	(3)
(Intercept)	0.0874***	0.4799***	0.1133***
	(0.0044)	(0.0070)	(0.0144)
prop65plus	**0.0039*** **		0.0036***
	(0.0001)		(0.0002)
prop0to14	0.0003	-0.0036***	
	(0.0002)	(0.0002)	
prop15to64		**-0.0039*** **	-0.0003
		(0.0001)	(0.0002)
Num.Obs.	6190	6190	6190
R2	0.254	0.254	0.254
R2 Adj.	0.253	0.253	0.253

* p < 0.05, ** p < 0.01, *** p < 0.001

This phenomenon occurs because any of the three variables can be expressed as 100 minus the other two. For example, using model (1) from Table 2 as the benchmark, and substituting $prop_{65+}$ with $(100 - prop_{0\text{-}14} - prop_{15\text{-}64})$, the model can be rewritten as

$$\begin{aligned}\text{Elective admission rate} &= 0.0003prop_{0\text{-}14} + 0.0039prop_{65+}\\ &= 0.0003prop_{0\text{-}14} + 0.0039(100 - prop_{0\text{-}14} - prop_{15\text{-}64})\\ &= (0.0003 - 0.0039)prop_{0\text{-}14} - 0.0039prop_{15\text{-}64}\\ &= -0.0036prop_{0\text{-}14} - 0.0039prop_{15\text{-}64}\end{aligned}$$

which gives the coefficients in model (2). Similar linear relations can also be observed when there are more than three compositions. Neither the significance nor the magnitude of the estimated coefficients of different pairs of proportions, which contain the exact same information, is consistent. The coefficients do not represent the *ceteris paribus* effects of the proportions on the outcome variable, as it is impossible to change any one proportion without simultaneously changing at least one other proportion. They are therefore meaningless, and it would be incorrect to interpret these coefficients in any way (Hron et al., 2012).

If the age proportions are not the variables of interest but are simply used to control for the age structure in a larger model, the estimated coefficients on any other covariates would be the same regardless of which age proportion category is omitted, and the model fit would also be unaffected by the decision over which age category to omit. However, the estimated coefficients on these other covariates would be biased because the raw form of the proportion of patients in each age group does not properly account for age structure in a linear regression. Including only one of the proportions as an explanatory variable would also result in incorrect estimates as this is essentially the same as having two proportions but arbitrarily omitting one from the model, which does not solve the non-linearity issue. As in 3 or more-part cases, the proportion of one of the two parts does not carry absolute information and using it would still violates the assumptions.

Note that the problem here is fundamentally different from the so-called “dummy variable trap”, i.e., including n dummy variables in the model to represent n mutually exclusive and exhaustive categories, which causes perfect collinearity (Wooldridge, 2015). To address the perfect collinearity issue in that instance one simply needs to omit one of the categories, the benchmark, and include only $n-1$ dummy variables. The coefficients of these dummy variables then represent the differences in the intercept of each group compared to that of the benchmark group, which always has an (omitted) dummy of 1. All other dummies can be zero and changing any one of them from 0 to 1 does not require the change of another dummy. This is not the case for compositional variables because any change to one of the components is necessarily accompanied by opposite changes of equal total value in the other components, i.e.,

$$\Delta x_j = -\sum_{i=1, i\neq j}^{D} \Delta x_i .$$

More importantly, the issues with using compositional variables in regressions does not stem from collinearity, but instead from the fact that they are co-dependent and only carry relative information.

Isometric log-ratio (ilr) transformation

Although the raw form of compositional data should not be used in traditional regression analysis, there are several ways to transform the compositions based on their log-ratios to allow for conventional statistical methods (Filzmoser et al, 2018). The pairwise ratios of our example dataset are displayed as boxplots in Figure 5.

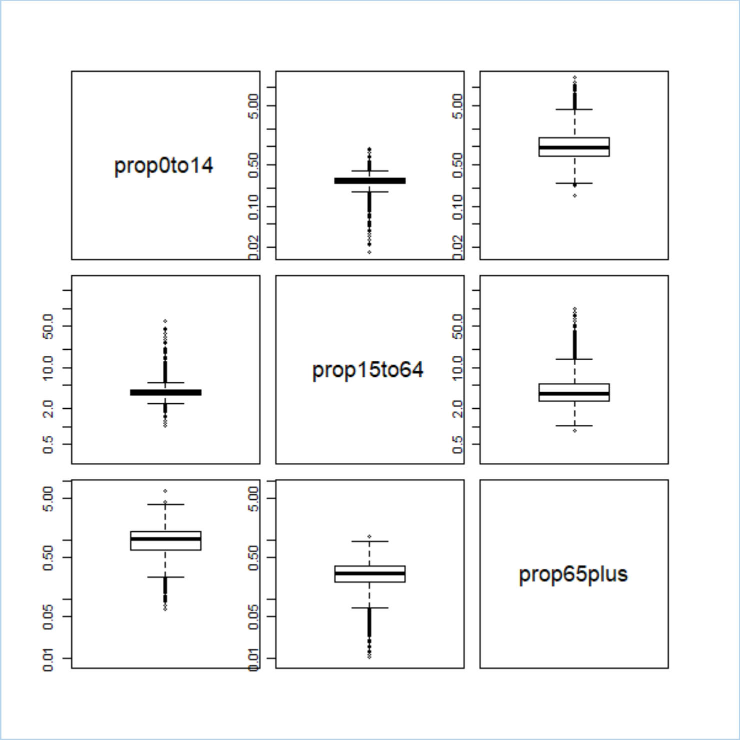

Figure 5 boxplots of pairwise log-ratios

Outliers with any of the three proportions lower than 1% are removed, as they lead to extremely high or low log-ratios and change the scale of the boxplots. Each boxplot shows the spread of the ratio between the corresponding two variables. The boxplots are put in the three-by-three table of plots to emphasize that they are "pairwise" but individual boxplot can be interpreted as usual. For example, the bottom left corner shows the spread of $\frac{prop_{0\text{-}14}}{prop_{65+}}$. Note that observations with the lowest pair-wise ratios do not necessarily have the lowest numerator proportions or highest denominator proportions.

The transformation method used in this paper is isometric log-ratio transformation, first introduced in Egozcue et al (2003). The key property of isometric log-ratio transformation is that it maps the data of d proportions from the simplex space $\mathcal{S}^d$ to the real space of $(d-1)$ dimensions, $\mathcal{R}^{d-1}$, keeping the relative positions between data points (Chastin and Palarea-Albaladejo, 2015). Other log-ratio methods such as additive log-ratio (alr) or centred log-ratio (clr) transformation can also achieve the same purpose, i.e., transforming the proportions to the Euclidean space in which linear regression in based. In general, isometric log-ratio has arguably the most desirable properties because the use of additive log-ratio does not preserve distances between data points due to its additive nature, while centred log-ratio transformed data are perfect collinear and require extra procedures to be used in linear regression (Aitchison, 1986; Hron et al., 2012; Pawlowsky-Glahn et al, 2015). Note that the log-ratio transformations are not the same as the logarithmic transformation ($z_i = \ln x_i, i = 1, \ldots, d$), which in this case does nothing but to change the original constraint to an equivalent $\sum_{i=1}^{d} e^{z_i} = 1$.

There are technically an infinite number of transformations that create isometry between the simplex and real space. A particularly versatile isometric transformation that can be used in different contexts is defined as follows (Egozcue et al, 2003; Hron et al., 2012):

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

Applying it to our example, we can obtain:

$$z_1 = \sqrt{\frac{3-1}{3-1+1}} \cdot \ln \frac{prop_{0\text{-}14}}{\sqrt[3-1]{prop_{15\text{-}64} \cdot prop_{65+}}} \tag{2}$$

$$= \sqrt{\frac{2}{3}} \ln \frac{prop_{0\text{-}14}}{\sqrt{prop_{15\text{-}64} \cdot prop_{65+}}};$$

$$z_2 = \sqrt{\frac{3-2}{3-2+1}} \cdot \ln \frac{prop_{15\text{-}64}}{prop_{65+}} \tag{3}$$

$$= \sqrt{\frac{1}{2}} \ln \frac{prop_{15\text{-}64}}{prop_{65+}}.$$

The proportions represent relative information which depend entirely on the ratios between the parts. Since the log-ratios between $prop_{0\text{-}14}$ and the other two proportions both appear in (2), z_1 contains the exact same information as $prop_{0\text{-}14}$. Similarly, z_2 does not have the exact same information as $prop_{15\text{-}64}$ because $prop_{0\text{-}14}$ does not appear in (3) and z_2 only includes the ratio between $prop_{15\text{-}64}$ and the remaining proportion, $prop_{65+}$.

However, each z_i uniquely represents the log-ratio of one pair of proportions, and this orthogonal projection from the simplex can guarantee consistency of distances and statistical analysis (Egozcue and Pawlowsky-Glahn, 2005). As shown by equation (1) , the order of the parts can be arbitrarily decided and the value or interpretation of z_1 is not affected by the order of the remaining parts. The regression results are also robust to the choice of z_1.

Using the transformed variables, we can estimate the following model:

$$\text{Elective admission rate} = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + e. \tag{4}$$

γ_1 indicates how much the ratios between $prop_{0\text{-}14}$ and other proportions is associated with the admission rate. The significance level can be interpreted as usual, i.e., whether the proportion explains the variation in admission rate, but it does not imply any independent effect of $prop_{0\text{-}14}$ on the admission rate. The regression result is shown in Table 3.

Table 3 Regression using raw and isometric log-transformed (ilr) data without covariates

	Raw		ilr
prop0to14	-0.0037***	z1	-0.0173***
	(0.0002)		(0.0033)
prop15to64	-0.0039***	z2	0.0593***
	(0.0001)		(0.0013)
(Intercept)	0.4799***	(Intercept)	0.2127***
	(0.0070)		(0.0034)
N	6190	N	6190
R2	0.254	R2	0.252
R2 Adj.	0.253	R2 Adj.	0.252

* p < 0.05, ** p < 0.01, *** p < 0.001

From equation (4) we can only draw conclusions about $prop_{0\text{-}14}$ as z_2 does not contain all of the ratios. To obtain the same information for $prop_{15\text{-}64}$ and $prop_{65+}$, we need to apply equation (1) to each of them and repeat the above process, as shown in Table 4. The permutation can be conducted manually or using R package "`robCompositions`" (Templ et al., 2011). The coefficients of z_2's in Table 4 should not be interpreted as they do not carry complete information. R functions such as "`lmCoDaX`" would report the estimators of the z_1 in each column as the coefficients of each proportion. If we are only using the proportions as control variables for the age structure in a model such as:

$$\text{Elective admission rate} = \boldsymbol{\gamma z} + \boldsymbol{\beta X} + e \tag{5}$$

where $\boldsymbol{X}$ contains income deprivation index, binary indicator of rurality, and time since last seen a GP, then no permutation is required because the coefficients for the covariates and the intercept would be the same. Note that we are only using this highly simplified model (5) to demonstrate the use of basic compositional data analysis techniques, and the coefficients in (5) inevitably suffer from omitted variable bias.

Table 4 Permutated regression results using each proportion as the first component

	$x_1 = prop_{0\text{-}14}$	$x_1 = prop_{15-65}$	$x_1 = prop_{65+}$
(Intercept)	0.1979***	0.1979***	0.1979***
	(0.0060)	(0.0060)	(0.0060)
z1	**-0.0065**	**-0.0573*** **	**0.0638*** **
	(0.0033)	**(0.0021)**	**(0.0021)**
z2	0.0699***	-0.0406***	-0.0293***
	(0.0015)	(0.0030)	(0.0030)
Income deprivation	0.1323***	0.1323***	0.1323***
	(0.0099)	(0.0099)	(0.0099)
Rurality	-0.0035*	-0.0035*	-0.0035*
	(0.0017)	(0.0017)	(0.0017)
Time since GP visit	-0.0013	-0.0013	-0.0013
	(0.0009)	(0.0009)	(0.0009)
Num.Obs.	6190	6190	6190
R2	0.277	0.277	0.277
R2 Adj.	0.276	0.276	0.276

* p < 0.05, ** p < 0.01, *** p < 0.001

Note that the coefficients do not represent the effect of a unit change in the explanatory variables on the admission rate. Rather, the effects of age group proportions should be considered as effects of log-ratios between each group, which can be quantified using a change prediction matrix as shown in Table 5 (Chastin and Palarea-Albaladejo, 2015). To calculate this matrix, we first apply inverse transformation to any pair of coefficients shown in Table 4 to get a composition vector:

$$W = ilrInv(-0.0065, 0.0699) = (0.3252, 0.3222, 0.3526).$$

The off-diagonal elements of the matrix are calculated as

$$m_{i,j} = \ln\left(\frac{w_i}{w_j}\right)/4d^2$$

where d is the number of parts which is equal to three in our data. This value represents how much the elective hospital admission rate reacts to a unit change of the log-ratio, i.e., $\ln\left(\frac{prop_i}{prop_j}\right)$. For example, column 1 row 3 of

Table 5 shows that if the log-ratio between $prop_{65+}$ and $prop_{0\text{-}14}$ increases by 1, equivalent to an increase in their ratio by the Euler's number $e \approx 2.72$,

$$\Delta \ln\left(\frac{prop_{65+}}{prop_{0\text{-}14}}\right) = 1 \Leftrightarrow \Delta \frac{prop_{65+}}{prop_{0\text{-}14}} = e$$

the admission rate is estimated to decrease by 0.0011.

Table 5 Change prediction matrix

	prop0to14	prop15to64	prop65+
prop0to14	0	0.0003	0.0011
prop15to64	-0.0003	0	0. 009
prop65+	-0.0011	-0.0009	0

Finally, we compare the fitted values of the model which uses the raw form of age group proportions and equation (5), which uses the isometric log-ratio transformations. It is important to note that both the values and the significance levels of the coefficients on the non-compositional variables of interest are different. The fitted values from these two models are plotted against the actual admission rates in Figure 6, and the difference between the fitted values is plotted against admission rates and the three age proportion variables in Figure 7. There appear to be a non-linear relationship between the difference and the proportions.

In practice, the difference caused by transforming the compositional data, or rather, the difference caused by incorrectly applying traditional regression analysis to compositional data, will vary by situation. However, the only way to determine the magnitude of the impact is to apply the transformations and then compare. More substantial differences between results obtained using raw and isometric log-ratio transformed were found for datasets with higher variations in the ratios between parts (Gupta et al., 2018). They also found that “CoDA and standard analyses may lead to different conclusions, not only from a numeric or statistical viewpoint, but even in terms of the practical applications of study results”.

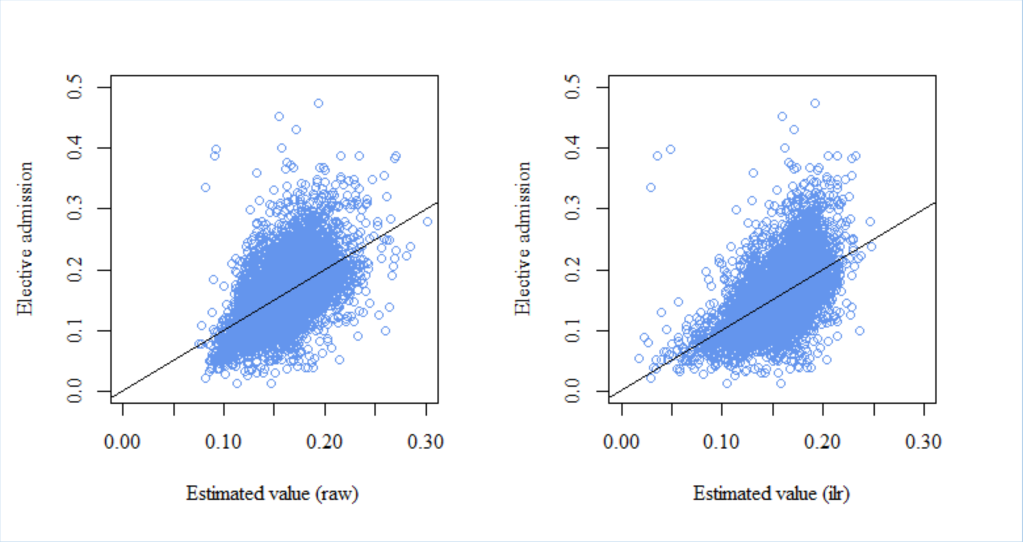

Figure 6 Actual elective admission rates vs fitted values using raw and transformed proportions

Table 6 Regression using raw and ilr-transformed data with covariates

	Raw		ilr
(Intercept)	0.0797***	(intercept)	0.1979***
	(0.0066)		(0.0060)
prop0to14	-0.0005*		
	(0.0002)		
prop15to64	0.0046***		
	(0.0001)		
Income deprivation	0.1341***	Income deprivation	0.1323***
	(0.0099)		(0.0099)
Rurality	-0.0096***	Rurality	-0.0035*
	(0.0017)		(0.0017)
Time since GP visit	-0.0016	Time since GP visit	-0.0013
	(0.0009)		(0.0009)
		z1	-0.0065+
			(0.0033)
		z2	0.0699***
			(0.0015)
Num.Obs.	6190		6190
R2	0.284		0.277
R2 Adj.	0.283		0.276

* p < 0.05, ** p < 0.01, *** p < 0.001

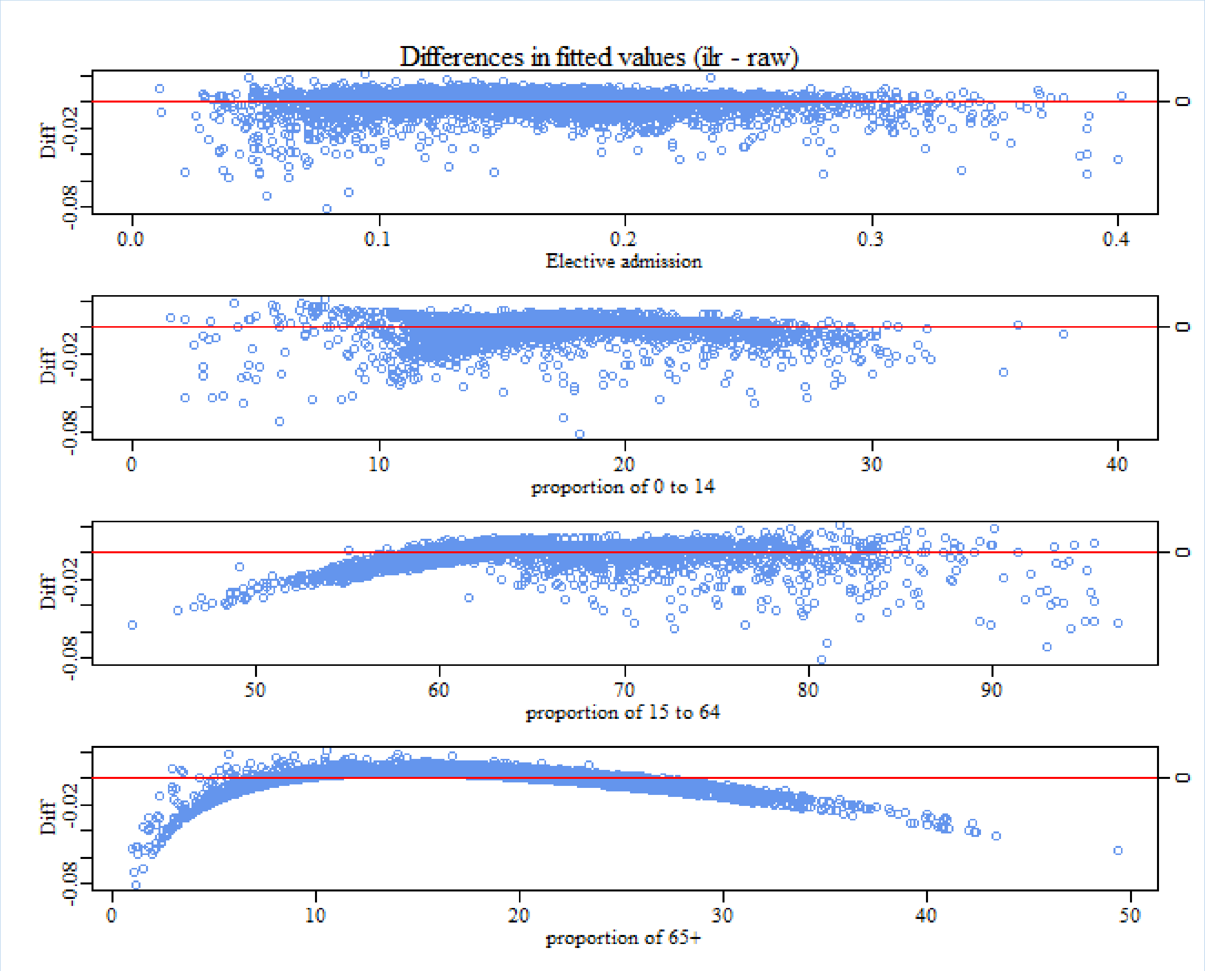

Figure 7 Differences in fitted values of models using transformed and raw proportion data

Composition of two parts

In this section we show that the issues discussed above also apply to compositions of two parts, and including only one in the linear regression does not produce reliable results as this approach does not address the codependence between the included and omitted variables caused by the constant sum. Using the same data, we now merge $prop_{0-14}$ and $prop_{15-64}$ into $prop_{under65}$. Then we repeat the steps above to compare the regression estimated using raw and ilr-transformed data, with the inclusion of the control variables. The results are shown in Table 7. After applying the transformation, the coefficient of rurality is no longer significant. The fitted values from are plotted against the actual elective admission rates in Figure 8, and the difference between the fitted values is plotted against admission rates and the 2 age group proportions (under 65, 65 and above) in Figure 8. As in the three-part case, there appears to be a non-linear relationship between the difference and the proportions.

Table 7 Regression using raw and ilr-transformed data with 1 proportion and covariates

	Raw		ilr
(Intercept)	0.5411***	(intercept)	0.2402***
	(0.0096)		(0.0054)
under65	-0.0047***		
	(0.0001)		
		z1	0.0783***
			(0.0017)
Income deprivation	0.1274***	Income deprivation	0.1167***
	(0.0095)		(0.0095)
Rurality	-0.0098***	Rurality	-0.0031
	(0.0017)		(0.0017)
Time since GP visit	-0.0014	Time since GP visit	-0.0008
	(0.0009)		(0.0009)
Num.Obs.	6190		6190
R2	0.283		0.274
R2 Adj.	0.282		0.273

* p < 0.05, ** p < 0.01, *** p < 0.001

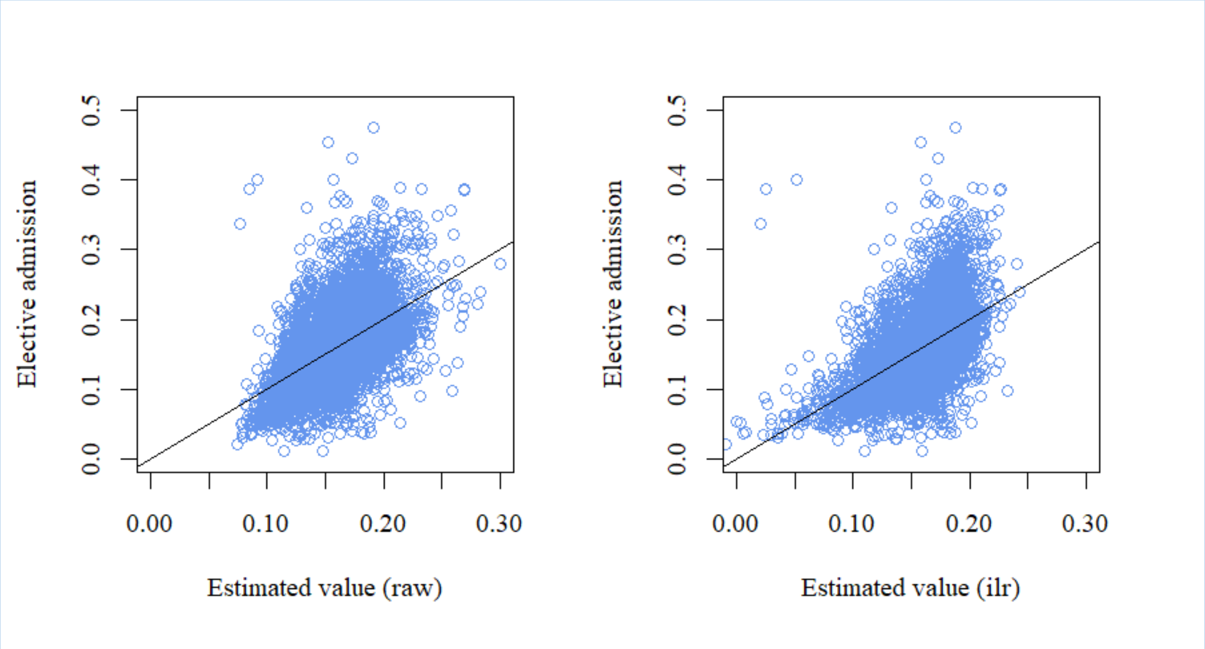

Figure 7 Actual elective admission rates vs fitted values using raw and transformed proportion of under 65

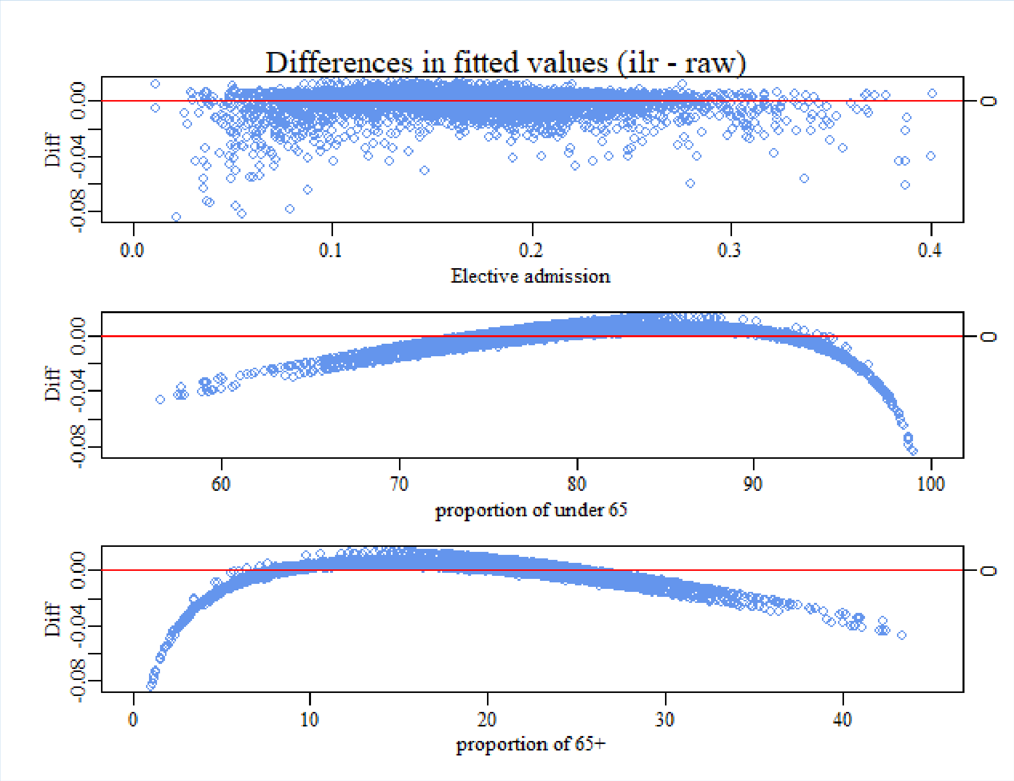

Figure 8 Differences in fitted values of models with one proportion, raw and transformed (four outliers with admission rate over 0.4 are removed)

Discussion

Compositional variables have special geometric properties, and applying standard statistical methods including regression analysis to the raw form of compositional data can produce misleading results. While compositional data analysis (CoDA) methods have been adopted in many research areas (Chastin et al, 2015; Greenacre, 2018; Mert et al., 2018), they are rarely seen in health economic research. This paper uses age group proportions as an example compositional variable to illustrate issues with compositional data analysis using traditional statistical techniques and provides a brief guide to basic CoDA methods, which can be applied to a wide range of variables that may be of interest to health economists, such as healthcare staff ratios and the proportion of resources allocated to health-related goods or activities.

While some of these variables, e.g., age proportions or racial composition, are often used as controls instead of being the primary variables of interest, health economists should nevertheless be aware that proportions follow different geometry to which conventional statistical methods are not suited. Directly using proportions in linear regression models does not properly account for their overall effect on the outcome, and the regression coefficients of the other variables in the model would also not be reliable as a result of their inclusion. Any basic log-ratio transformation is sufficient to avoid the issues.

The transformations do not necessarily lead to a higher R^2 of the regression model but implementing them reduces the risk of observing spurious effects and ensures robustness. Researchers should be cautious analysing data of compositional nature and use appropriate CoDA methods, otherwise it is better to avoid the use of compositional data altogether. Analysing the absolute numbers underlying the proportions (e.g. number of patients in each age band) is an option if available, although they would often be highly correlated with the counts of other variables, (e.g. the number of medical staff in the area). Using the underlying counts is not an option for variables such as proportion of time spent on different activity during a day, as the raw values also have a constant sum.

A variety of R packages have been developed in recent years to allow convenient application of compositional data transformations and to provide sensible interpretations (Filzmoser et al., 2018; Quinn et al., 2017). We are currently working on developing an equivalent Stata package. As this paper only aims to introduce the concept of compositional data and explain some basic analyses, the existence of irregular data including zeroes and outliers is not discussed, however it is covered in many texts (Pawlowsky-Glahn and Buccianti, 2011; Van den Boogaart and Tolosana-Delgado, 2013).

References

AITCHISON, J. 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological),* 44**,** 139-160.

AITCHISON, J. 2008. The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies.

ARNOLD, K. F., BERRIE, L., TENNANT, P. W. G. & GILTHORPE, M. S. 2020. A causal inference perspective on the analysis of compositional data. *International Journal of Epidemiology,* 49**,** 1307-1313.

BUCCIANTI, A., MATEU-FIGUERAS, G. & PAWLOWSKY-GLAHN, V. Compositional data analysis in the geosciences: from theory to practice. 2006. Geological Society of London.

CHASTIN, S. & PALAREA-ALBALADEJO, J. 2015. Concise guide to compositional data analysis for physical activity, sedentary behaviour and sleep research: supplementary material S2. *Chastin SFM, Palarea-Albaladejo J, Dontje ML, Skelton DA. Combined effects of time spent in physical activity, sede. PLoS One,* 10**,** e0139984.

CHASTIN, S. F., PALAREA-ALBALADEJO, J., DONTJE, M. L. & SKELTON, D. A. 2015. Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. *PloS one,* 10**,** e0139984.

CHAYES, F. 1960. On correlation between variables of constant sum. *Journal of Geophysical research,* 65**,** 4185-4193.

CHAYES, F. 1962. Numerical correlation and petrographic variation. *The Journal of Geology,* 70**,** 440-452.

DRASTICHOVÁ, M. & FILZMOSER, P. 2020. The relationship between health outcomes and health expenditure in Europe by using compositional data analysis. *Problemy Ekorozwoju,* 15**,** 99-110.

EGOZCUE, J. J. & PAWLOWSKY-GLAHN, V. 2005. Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology,* 37**,** 795-828.

EGOZCUE, J. J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. & BARCELÓ-VIDAL, C. 2003. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology,* 35**,** 279-300.

FILZMOSER, P., HRON, K. & TEMPL, M. 2018. Applied compositional data analysis. *Cham: Springer*.

GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. & EGOZCUE, J. J. 2017. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology,* 8**,** 2224.

GREENACRE, M. 2018. *Compositional data analysis in practice*, CRC press.

GUPTA, N., MATHIASSEN, S. E., MATEU-FIGUERAS, G., HEIDEN, M., HALLMAN, D. M., JØRGENSEN, M. B. & HOLTERMANN, A. 2018. A comparison of standard and compositional data analysis in studies addressing group differences in sedentary behavior and physical activity. *International Journal of Behavioral Nutrition and Physical Activity,* 15**,** 53.

HRON, K., FILZMOSER, P. & THOMPSON, K. 2012. Linear regression with compositional explanatory variables. *Journal of applied statistics,* 39**,** 1115-1128.

MERT, M. C., FILZMOSER, P., ENDEL, G. & WILBACHER, I. 2018. Compositional data analysis in epidemiology. *Statistical Methods in Medical Research,* 27**,** 1878-1891.

NHS-DIGITAL 2020. NHS Payments to General Practice - England, 2019/20

NHS-DIGITAL 2023. General Practice Workforce. *In:* NHS-DIGITAL (ed.).

PAWLOWSKY-GLAHN, V. & BUCCIANTI, A. 2011. *Compositional data analysis*, Wiley Online Library.

PAWLOWSKY-GLAHN, V. & EGOZCUE, J. J. 2001. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment,* 15**,** 384-398.

PAWLOWSKY-GLAHN, V. & EGOZCUE, J. J. 2006. Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications,* 264**,** 1-10.

PAWLOWSKY-GLAHN, V., EGOZCUE, J. J. & TOLOSANA-DELGADO, R. 2015. *Modeling and analysis of compositional data*, John Wiley & Sons.

PEARSON, K. 1897. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london,* 60**,** 489-498.

QUINN, T. P., RICHARDSON, M. F., LOVELL, D. & CROWLEY, T. M. 2017. propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Scientific reports,* 7**,** 1-9.

ROCK, N. M. 1988. Section V: Some Special types of geological data. *Numerical Geology: A Source Guide, Glossary and Selective Bibliography to Geological Uses of Computers and Statistics***,** 200-245.

TEMPL, M., HRON, K. & FILZMOSER, P. 2011. robCompositions: an R-package for robust statistical analysis of compositional data. *Compositional data analysis: Theory and applications***,** 341-355.

URWIN, S., LAU, Y. S., GRANDE, G. & SUTTON, M. 2023. Informal caregiving, time use and experienced wellbeing. *Health Economics*.

VAN DEN BOOGAART, K. G., TOLOSANA, R., BREN, M. & VAN DEN BOOGAART, M. K. G. 2013. Package ‘compositions’. *Compositional data analysis. Ver***,** 1-40.

VAN DEN BOOGAART, K. G. & TOLOSANA-DELGADO, R. 2013. *Analyzing compositional data with R,* Springer.

VISTELIUS, A. B. & SARMANOV, O. V. 1961. On the Correlation between Percentage Values: Major Component Correlation in Ferromagnesium Micas. *The Journal of Geology,* 69**,** 145-153.

WOOLDRIDGE, J. M. 2015. *Introductory econometrics: A modern approach,* Cengage learning.