Citation for published version:
Dumuid, D, Stanford, TE, Martin-Fernández, JA, Pedisic, Z, Maher, C, Lewis, L, Hron, K, Katzmarzyk, PT, Chaput, J-P, Fogelholm, M, Hu, G, Lambert, EV, Maia, J, Sarmiento, OL, Standage, M, Barreira, TV, Broyles, ST, Tudor-Locke, C, Tremblay, MS & Olds, T 2018, 'Compositional data analysis for physical activity, sedentary time and sleep research', *Statistical Methods in Medical Research*, vol. 27, no. 12, pp. 3726-3738. https://doi.org/10.1177/0962280217710835

DOI:
10.1177/0962280217710835

Publication date:
2018

Document Version
Peer reviewed version

Link to publication

Dorothea Dumuid, Tyman E Stanford, Josep-Antoni Martin-Fernández, Željko Pedišić, Carol A Maher, Lucy K Lewis, Karel Hron, Peter T Katzmarzyk, Jean-Philippe Chaput, Mikael Fogelholm, Gang Hu, Estelle V Lambert, José Maia, Olga L Sarmiento, Martyn Standage, Tiago V Barreira, Stephanie T Broyles, Catrine Tudor-Locke, Mark S Tremblay, Timothy Olds, Compositional data analysis for physical activity, sedentary time and sleep research, Statistical Methods in Medical Research. Copyright © 2017 The Author(s). Reprinted by permission of SAGE Publications.

University of Bath

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 04. Sep. 2026

1 Introduction: Daily activity behaviours are compositional by nature

Physical inactivity is considered to be a major risk factor for non-communicable disease and premature death.[1] A global economic analysis has estimated the health-care system cost of physical inactivity in 2013 alone to be $ (INT$) 53.8 billion.[2] To produce such economic estimates, physical inactivity is defined as relatively lower amounts of moderate-to-vigorous-intensity physical activity (MVPA), but underlying analyses fail to account for the fact that when time in MVPA is reduced, a subsequent and equal increase in time must be distributed to the remaining behaviour domains: sleep, sedentary time and light-intensity physical activity (light PA), to represent the finite 24-hours, or 1440 minutes, in any given day. These other non-MVPA behaviours may themselves have positive or negative effects on health and mortality.[3-6] Therefore, the health and economic burden of physical inactivity per se remains unclear. Similar estimates for sedentary time [7] are uncertain for the same reason, i.e., they fail to adequately account for other behaviours. Pedišić [8] argued that studies on health outcomes of sleep duration and light PA can be put under the same scrutiny.

Time spent in MVPA represents one of the exhaustive and mutually exclusive components of an individual's 24-h day. The non-MVPA time remaining within an individual's day can be partitioned into light PA, sedentary time and sleep and all of

them can be considered as relative contributions to the overall time budget. The defined behaviours (MVPA, light PA, sedentary time and sleep) are therefore compositional data, and have important properties that must be respected.[9]

Consider a vector $\boldsymbol{x} = [x_1, x_2, \dots, x_D] \in \mathbb{R}^D$ with positive components, where $\sum x_i = C$, and C is the closure constant. The sample space of the vector $\boldsymbol{x}$ can thus be represented by the D-part simplex ($\boldsymbol{S}^{\mathrm{D}}$), which is a ($D$-1)-dimensional subset of the real space ($\boldsymbol{R}^{\mathrm{D}}$) due to the constant sum constraint of C. Compositional data are scale invariant[9] because the application of a common factor a to the parts x_i where $i = 1,2, \dots, D$ ensures the relative difference between the parts is maintained, as $\sum_{i=1}^{D} ax_i = a\sum_{i=1}^{D} x_i = aC$. The numerical value of the closure constant (e.g., 24 h, one week, one month) is irrelevant. Daily behaviours could equivalently be measured in hours, minutes or percentages as the data convey only relative information. The property of scale invariance means compositional data are in fact elements of equivalence classes of proportional vectors.[10] Accordingly, the simplex is the sample space of representatives of compositional data with a chosen constant sum constraint. Specific properties of compositional data are followed by a natural geometry, known as the *Aitchison geometry*.[11] The closure constant representation of compositions imposes perfect multi-collinearity among the components, causing the covariance structure of the data to be negatively biased.[9] Accordingly, traditional statistical methods for unconstrained variables in real space

(e.g., t-tests, multiple linear regression) are not predicative with respect to the specific geometry of the simplex sample space, and should not be used for absolute or raw measures of time spent in daily behaviours.[12]

2 The log-ratio approach for compositional data analysis

The invalidity of standard multivariate techniques for analyzing untransformed or raw compositional data was recognized in scientific fields decades ago,[13] and in the 1980s Aitchison proposed a new methodology for the analysis of compositional data.[9] The methodology is based on the premise that any composition (e.g., an individual's daily time budget) can be expressed in terms of ratios of its parts (e.g., duration of sleep, sedentary time, light PA and MVPA). The expression of compositional data as log-ratio coordinates transfers them from the constrained simplex space to the unconstrained real space, where traditional multivariate statistics may be applied.[14] The presence of zeros in a compositional dataset prohibits applying log-ratio coordinates. Several methods have been proposed to deal with zeros;[15] however they are beyond the scope of this paper.

A number of log-ratio coordinate systems for compositional data have been described.[11] One such coordinate system, the additive log-ratio (*alr*), has coordinates, $\boldsymbol{a}$, defined by

$$\boldsymbol{a} = [a_1, \dots, a_{D-1}] = alr(\boldsymbol{x}) = [\ln\left(\frac{x_1}{x_D}\right), \ln\left(\frac{x_2}{x_D}\right), \dots, \ln\left(\frac{x_{D-1}}{x_D}\right)]. \quad (1)$$

However, the *alr* coordinates are asymmetric, because the components $x_1, x_2, \ldots, x_{D-1}$ are divided by the component x_D. Moreover, they are not isometric, i.e., distances and angles in the Aitchison geometry are violated by using the *alr* coordinates, limiting their use in statistical applications. This means that the system of *alr* coordinates in the Aitchison geometry is oblique, and traditional statistical methods which assume orthogonal coordinates therefore cannot be directly applied. Another coordinate system is the centred log-ratio (*clr)* coordinate system, $\boldsymbol{c}$, defined as

$$\boldsymbol{c} = [c_1, \ldots, c_D] = clr(\boldsymbol{x}) = \left[\ln\left(\frac{x_1}{\tilde{x}}\right), \ln\left(\frac{x_2}{\tilde{x}}\right), \ldots, \ln\left(\frac{x_D}{\tilde{x}}\right)\right], \quad (2)$$

where $\tilde{x}$ is the geometric mean of all the D components of the vector $\boldsymbol{x}$. The *clr* are symmetric and isometric; they produce a singular covariance matrix because $\sum_{j=1}^{D} c_j = 0$. The *clr* are, strictly speaking, not coordinates but coefficients with respect to a generating system. The covariance matrix of *clr* coefficients is singular, so the *clr* coefficients cannot be fully utilized as independent variables in multiple regression analysis.

The singularity problem of the *clr* can be overcome by the use of an isometric log-ratio (*ilr*) coordinate system. Isometric log ratio coordinates form an isometric mapping of the composition from the simplex sample space to the real space.[16] To construct the *ilr* coordinates, an orthonormal basis coherent with the Aitchison geometry is created in the (*D*-1)-dimensional hyperplane of the *clr* coordinates. Many possible orthonormal

coordinate systems can be created, however, following Pawlowsky-Glahn et al.,[11] one can define a particular *ilr* system of coordinates through a partitioning process.[16] For our purposes, specific *ilr* coordinates based on a sequential partition of one part to the remaining compositional parts[17] is very useful. Such *ilr* coordinates result in a (*D*-1)-dimensional real vector, **z,** defined as

$$\boldsymbol{z} = [z_1, z_2, \dots, z_{D-1}] = ilr(\boldsymbol{x}) = \left[\sqrt{\frac{D-1}{D}}\ln\left(\frac{x_1}{\sqrt[D-1]{\prod_{k=2}^{D} x_k}}\right), \sqrt{\frac{D-2}{D-1}}\ln\left(\frac{x_2}{\sqrt[D-2]{\prod_{k=3}^{D} x_k}}\right), \dots\right.$$

$$\left. \dots, \sqrt{\frac{D-j}{D-j+1}}\ln\left(\frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}\right), \dots, \frac{1}{\sqrt{2}}\ln\left(\frac{x_{D-1}}{x_D}\right)\right]. \quad (3)$$

It can be shown that

$$c_1 = \sqrt{\frac{D-1}{D}}\, z_1, \quad (4)$$

i.e., the first *ilr* coordinate in this system is directly proportional to the first *clr* coefficient. Both c_1 and z_1 can be interpreted in the same way in terms of dominance of a component (here the first compositional part) to the rest of parts. The remaining *ilr* coordinates (z_2, z_3, …,z_{D-1}) contain no relative information regarding the first compositional part. If a different compositional part is of interest, e.g., the second part, it is simply a matter of rearranging the compositional parts so that the part of interest is in the first place, and then reconstructing the *ilr* coordinates according to (3).[18]

The *ilr* coordinates can also be constructed using a sequential binary partition (SBP) as described by Egozcue and Pawlowsky-Glahn.[16] The first step in the SBP process requires division of the full composition into two subgroups of parts, where one subgroup will form the numerator and the other the denominator of the first *ilr* coordinate. In the subsequent steps, each of these two subgroups is further split into new subgroups to create the remaining *ilr* coordinates. For example, based on prior knowledge of the nature of the behaviour components, the numerator could be selected to consist of inactivity-related behaviours (sleep and sedentary time) and the denominator could be activity-related components (light PA and MVPA). The second SBP then divides the numerator of the first *ilr* coordinate to form the second *ilr* coordinate, with sleep as the numerator and sedentary time as the denominator. The final SBP divides the denominator of the first *ilr* coordinate, with light PA as the numerator and MVPA the denominator of the third *ilr* coordinate.

As the *ilr* coordinates defined in (3) are orthogonal, the columns of the design matrix are linearly independent. This avoids the previous issue of a singular covariance matrix in the multiple linear regression fit. The regression parameters estimated for the first *ilr* coordinate in the model represent the effect on an outcome when the first component (numerator) is changed in relation the geometric mean of the remaining parts (denominator). To examine the influence of the other compositional parts (relative to

the geometric mean of the respective remaining parts), a total of *D* models are fitted, with each model including a set of *ilr* which iteratively has a different compositional part as the numerator of the first *ilr* coordinate (and the remaining parts as the denominator). The constant term, and other external covariate terms, as well as the quality of fit, are invariant to the choice of *ilr* basis.[11]

3. Compositional data analysis in the field of sleep, sedentary and physical activity research

The log-ratio approach for compositional data analysis is well established in many scientific fields (e.g., geology, biology, hydrology, ecology and economics), and is considered the gold-standard for analyzing compositional data.[11] However, this methodology has only recently been used in health research, with researchers applying compositional data analysis to nutrition,[19] epidemiology[20] and microbiome data.[21] Furthermore, the compositional nature of daily activity data (sleep, sedentary time, light PA and MVPA) was not acknowledged until 2014, when Pedišić [8] warned that traditional analyses within the field were undermined due to the use of inappropriate and invalid statistical procedures, and called for a paradigm shift towards a compositional approach. Subsequently, Chastin et al. [18] and Carson et al. [22] demonstrated the feasibility of estimating the relationship between the complete daily behaviour composition and health outcomes using the *ilr* methodology outlined above.

However, the interpretation of log-ratio regression coefficients is not straight-forward.[23] As daily activity data have a meaningful total, i.e., 24-hours or 1440 minutes, regression coefficients can be interpreted on a meaningful scale.

The *ilr* multiple linear regression model for n compositional observations $(\boldsymbol{x_i}, y_i),\ i = 1,2,\dots,n,$ where $\boldsymbol{x_i} = [x_{i1}, x_{i2}, \dots, x_{iD}]$ with $\sum_{j=1}^{D} x_{ij} = 1$, is

$$y_i = \beta_0 + \sum_{j=1}^{D-1} \beta_j z_{ij} + \varepsilon_i \tag{5}$$

where

$$z_{ij} = \sqrt{\frac{D-j}{D-j+1}} \ln\left(\frac{x_{ij}}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_{ik}}}\right) \text{ for } j = 1,2,\dots,D-1,$$

with intercept β_0, regression parameters $\beta_1, \beta_2, \dots, \beta_{D-1}$ and $\varepsilon_i \sim N(0, \sigma^2)$ independently. The regression coefficient β_1 represents the change in the response variable when the first *ilr* coordinate is changed while the remaining *ilr* coordinates are all kept constant and the total sum is maintained, i.e., $\sum_{j=1}^{D} x_{ij} = 1$. When the component in the numerator of the first *ilr* coordinate is increased by one factor, all the components in the denominator can simultaneously be decreased by another factor to maintain the constant total. For example, ilr_1 for a 4-part composition becomes

$$\sqrt{\frac{3}{4}} \ln\left(\frac{x_1}{\sqrt[3]{(x_2 \cdot x_3 \cdot x_4)}} \cdot \frac{1+r}{1-s}\right). \tag{6}$$

The remaining *ilr* coordinates are then all kept constant, as they contain only the components from the denominator of the first *ilr* coordinate which are all decreased by the same proportion. Continuing with the example above, the remaining *ilr* coordinates, i.e., ilr_2 and ilr_3, respectively become

$$\sqrt{\frac{2}{3}}\ln\left(\frac{x_2}{\sqrt{(x_3 \cdot x_4)}} \cdot \frac{1-s}{1-s}\right) \text{ and } \sqrt{\frac{1}{2}}\ln\left(\frac{x_3}{x_4} \cdot \frac{1-s}{1-s}\right). \quad (7)$$

The estimated change in an outcome ($\triangle\hat{y}$) when the first component (x_1) of a composition of interest (e.g., the mean composition for a particular population group) is multiplied by 1+*r* (other coordinates are kept constant as each remaining compositional part is simultaneously multiplied by 1-*s*) can be calculated as

$$\triangle\hat{y} = \hat{\beta}_1 \cdot \sqrt{\frac{D-1}{D}} \cdot \ln\left(\frac{1+r}{1-s}\right) \quad (8)$$

$$\text{where } -1 < r < \frac{1-x_1}{x_1} \text{ and } s = r \cdot \frac{x_1}{1-x_1}.$$

The derivation of Equation (8) above, the corresponding 100(1-α)% confidence interval and the effect of including additional predictor variables into the *ilr* multiple linear regression model is provided in Supplementary file 1.

To facilitate the translation of research findings to clinical practice, it is of interest to estimate the health effects related to a meaningful quantum change (in minutes or hours) of one part of the activity behaviour composition (relative to compensatory change – to

maintain a total of 24 hours – of the geometric mean of the remaining compositional parts). The next section illustrates a novel, meaningful interpretation of *ilr* beta coefficients from regression analysis of a specific type of data: compositional, and constrained to a meaningful constant sum (24 h or 1440 min), using an epidemiological dataset as an example.

4 Example: daily activity and adiposity

The examples use data from the International Study of Childhood Obesity, Lifestyle and the Environment (ISCOLE), a large international study of children aged 9-11 years, conducted between 2011 and 2013.[24] Children were from urban and suburban centres in 12 countries (Australia, Brazil, Canada, China, Colombia, Finland, India, Kenya, Portugal, South Africa, England, and the United States). Ethical approval for ISCOLE was obtained from the Institutional Review Board of the Pennington Biomedical Research Center in Baton Rouge, Louisiana, USA, and site-specific ethical approval was also received at each participating study site. Parental written informed consent and child assent were obtained as required by local review boards. Daily activity was measured by 7-day 24-hour accelerometry.[25] Nocturnal sleep duration was estimated using a fully automated algorithm.[26, 27] Once total sleep time and awake non-wear time (any sequence of ≥20 consecutive minutes of 0 activity counts) were removed, data were processed in 15-s epochs to determine sedentary time (≤25 counts per 15 s), light

PA (26-573 counts per 15 s), and MVPA (≥574 counts per 15 s), congruent with Evenson's cut-points.[28] For analysis, each component of 24-h time use (sleep, sedentary time, light PA and MVPA) was accumulated, weighted for weekdays:weekend days at 5:2. No zero values were present among the compositional daily activity behaviour data. Adiposity was represented by body mass index (BMI) from measured weight and height (BMI= weight [kg]/height [m^2]). Note that BMI is a positive real random variable with an absolute scale. Participant BMI was converted to z-scores using age- and sex-specific World Health Organization (WHO) reference data.[29] After its transformation, zBMI can take on positive and negative values, satisfying the assumption that the response variable has support on the real line. The analyses included 5828 children (2633 boys, 3195 girls), with mean zBMI= 0.45 (SD=1.26).

First, the relationship between zBMI and the four-part daily activity composition (sleep, sedentary time, light PA, MVPA) was examined. Table 1 shows the results of the compositional multiple linear regression models, which included the *ilr* coordinates and terms for sex, highest parental education, number of parents, number of siblings, and study site. The *ilr* coordinates were calculated using (3) in which $\boldsymbol{x}$ represented proportions of time spent in sleep, sedentary time, light PA and MVPA. Four sets of *ilr*-coordinate systems were constructed, each time rotating the sequence of activity behaviours, so that each behaviour was iteratively represented as the first compositional

part. Each of these *ilr*-coordinate systems was used in a multiple linear regression model, where the regression coefficient of the first *ilr*-coordinate contained all the information regarding the first activity component, relative to all the remaining components. Therefore, only the regression coefficients corresponding to the first *ilr*-coordinates are displayed in Table 1.

Table 1. Multiple linear regression analyses of the relationship between first isometric log-ratio (*ilr*) coordinates and Body Mass Index (BMI) z-scores (Compositional models).

ilr regression models	$\hat{\beta}$	SE	*t*-value	*p*
Model 1: $ilr_1 \propto$ ln(Sleep : geometric mean of remaining behaviours)	-0.82	0.13	-6.22	<0.001
Model 2: $ilr_1 \propto$ ln(Sedentary time : geometric mean of remaining behaviours)	0.35	0.10	3.30	<0.001
Model 3: $ilr_1 \propto$ ln(Light PA :	1.34	0.10	13.19	<0.001

geometric mean of remaining behaviours)				
Model 4: $ilr_1 \propto$ ln(MVPA : geometric mean of remaining behaviours)	-0.87	0.05	-16.11	<0.001

ilr: Isometric log ratio, BMI z-score: Body Mass Index transformed to z-score using age- and sex-specific World Health Organization (WHO) reference data, $\hat{\beta}$: unstandardized regression coefficient estimate, SE: Standard error, Light PA: Light-intensity physical activity; MVPA: Moderate-to-vigorous-intensity physical activity. All models adjusted for sex, highest parental education level, number of siblings, number of parents and study site. Adjusted *R*-squared = 0.11.

Adiposity (zBMI) was positively related to the relative time spent in sedentary time and light PA, and negatively related to the relative time spent in sleep and MVPA. Figure 1 shows the association effect-size from the linear models further interpreted to meaningful parameters. To create the plot, the estimated differences in zBMI related to difference in one activity relative to the remaining activities (with the mean activity behaviour composition of the sample as the reference, or starting composition) were calculated using (8) for 15 min time interval increments ranging from 0 to 60 min

(MVPA) or 0 to 120 minutes (sleep, sedentary time, light PA). Worked examples are presented in Supplementary file 2. The R functions to estimate differences in zBMI are available freely at https://github.com/tystan/deltacomp. From Figure 1, it can be seen that estimated zBMI is lower by about 0.33 units with 30 min higher relative MVPA, and zBMI is higher by about 0.60 units with 30 min lower relative MVPA, compared to zBMI at the mean composition. The non-linear/non-symmetrical nature of the estimated zBMI response curves can be seen in Figure 1. The figure can also be used to assess equivalence of activity behaviours in the relationship with zBMI; for example, a horizontal line drawn at a zBMI -0.1 shows that this estimated difference in zBMI is associated with either 48 min more sleep (relative to the remaining behaviours), 108 min less sedentary time (relative to the remaining behaviours), 21 min less light PA (relative to the remaining behaviours) or 8 min more MVPA (relative to the remaining behaviours) than the mean activity behaviour composition. The minute values can be calculated from the linear models, as detailed in Supplementary file 2.

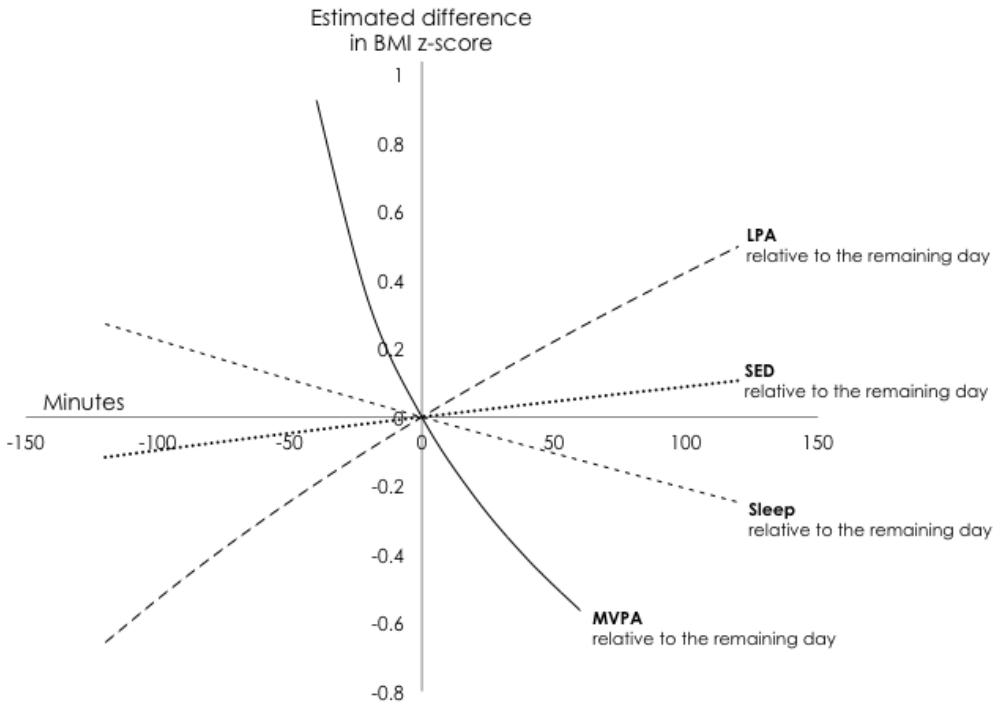


Figure 1. The Relationship Between Daily Behaviours and zBMI, Estimated by Compositional Linear Regression Models.

BMI: Body Mass Index; SED: Sedentary Time; LPA: Light-Intensity Physical Activity; MVPA: Moderate-to-Vigorous-Intensity Physical Activity.

Difference in Minutes Modelled Around the Population Mean Composition of (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45.

The relationship between zBMI and the activity behaviour composition was further investigated using an SBP approach, with *ilr* coordinates constructed according to procedures outlined in Egozcue & Pawlowsky-Glahn.[16] The first partition separated the behaviour components into two groups; inactivity-related behaviours (sleep and sedentary time), and activity-related behaviours (light PA and MVPA). The second partition was between sleep and sedentary time, and the final partition was between light PA and MVPA. Table 2 presents the results from the *ilr* log-ratio multiple linear regression model based on this SBP.

Table 2. Multiple linear regression analysis of the relationship between the isometric log-ratio (*ilr*) coordinates obtained from a sequential binary partition and Body Mass Index (BMI) z-scores (Compositional model).

	$\hat{\beta}$	SE	*t*-value	*p*
$ilr_1 \propto$ ln(geometric mean of Sleep & Sedentary time : geometric mean of Light PA & MVPA)	-0.41	0.08	-5.33	<0.001
$ilr_2 \propto$ ln(Sleep : Sedentary time)	-0.72	0.13	-5.29	<0.001

$ilr_3 \propto$ ln(Light PA : MVPA)	-1.35	0.08	-16.26	<0.001

ilr: Isometric log ratio, BMI z-score: Body Mass Index transformed to z-score using age- and sex-specific World Health Organization (WHO) reference data, $\hat{\beta}$: unstandardized regression coefficient estimate, SE: Standard error, Light PA: Light-intensity physical activity, MVPA: Moderate-to-vigorous intensity physical activity. Model adjusted for sex, highest parental education level, number of siblings, number of parents and study site. Adjusted *R*-squared = 0.11.

Regression parameters suggest that, with increase in inactivity-related behaviours relative to decrease in activity-related behaviours, zBMI decreases (Table 2). This finding is explained by the co-consideration of light PA and MVPA as activity-related behaviours, i.e., the denominator of the first *ilr* coordinate is the geometric mean of both light PA and MVPA. As shown in the previous analysis (Table 1), light PA (relative to all remaining behaviours) has a strong positive association with zBMI. Predicted increase in zBMI for increase in light PA (accompanied by corresponding decrease in all other behaviours) was higher than the respective effect sizes for any other behaviour (relative to the remaining behaviours) (Figure 1). This is because whatever light PA replaces (sleep, MVPA, and more surprisingly, sedentary time) is associated with lower fatness. The regression coefficient for the second *ilr* coordinate from the SBP (Table 2) implies that the increase in sleep relative to sedentary time is associated with lower

expected zBMI. The third *ilr* regression coefficient indicates that an increase in light PA at the expense of MVPA is associated with higher expected zBMI. The results obtained from the regression model from the SBP are consistent and complementary to the results from the previous *ilr* models. With the SBP we obtain complementary information about the substitution of time between selected groups of parts. Of course, with an SBP as presented here, none of the coordinates extract all the relative information about any of the behaviours. This might be an advantage because of the danger that the geometric mean of the other components may itself conceal some unpredictable patterns and potentially affect the interpretability of the first coordinate. On the other hand, the concrete choice of SBP necessarily entails a substantial amount of subjectivity and, additionally, the effects of merging information contained in compositional parts by taking the geometric mean will also influence the interpretation of the first coordinate.

To further illustrate the importance of using a compositional approach, we analyzed the same dataset using standard (non-compositional) regression (Table 3). The purpose of ensuing analysis is not to compare the findings of standard regression models to findings of compositional regression models, but to demonstrate the potential pitfalls and limitations of standard multiple linear regression when analyzing data of a compositional nature. Raw values of time spent in each behaviour (min/day) were used,

with zBMI as the dependent variable. It is not possible to include all daily activity behaviours (sleep, sedentary time, light PA and MVPA) in the regression model as this would result in a singular covariance matrix. Therefore four models were used, with each model iteratively excluding a different behaviour, i.e., (1) excluded sleep; (2) excluded sedentary time; (3) excluded light PA; (4) excluded MVPA. The four analyzed models represent the most adjusted traditional regression models. Variance Inflation Factors (VIF) for Models 1 - 3 ranged between 1.9 and 2.4, indicating multi-collinearity was likely not a concern.[30] However, VIF for model 4 were high (between 7.2 and 11.3), suggesting potential instability of regression estimates. Each model was additionally adjusted for sex, highest parental education, number of parents, number of siblings, and study site.

Table 3. Traditional multiple linear regression analysis of the relationship between raw daily activity data (min/day) and Body Mass Index (BMI) z-scores (Non-compositional models).

	$\hat{\beta}$	SE	*t*-value	*p*
Model 1				
(Sleep excluded)				
Sedentary time	0.002	0.000	5.27	<0.001
Light PA	0.005	0.000	11.48	<0.001

MVPA	-0.011	0.001	-13.17	<0.001
Model 2				
Sleep (Sedentary time excluded)	-0.002	0.000	-5.27	<0.001
Light PA	0.003	0.000	8.33	<0.001
MVPA	-0.013	0.001	-17.32	<0.001
Model 3				
Sleep	-0.005	0.000	-11.48	<0.001
Sedentary time (Light PA excluded)	-0.003	0.000	-8.33	<0.001
MVPA	-0.016	0.001	-17.35	<0.001
Model 4				
Sleep	0.011	0.001	13.17	<0.001
Sedentary time	0.013	0.001	17.32	<0.001

Light PA (MVPA excluded)	0.016	0.001	17.35	<0.001

BMI z-score: Body Mass Index transformed to z-score using age- and sex-specific World Health Organization (WHO) reference data, $\hat{\beta}$: unstandardized regression coefficient estimate, SE: Standard error, Light PA: Light-intensity physical activity, MVPA: Moderate-to-vigorous-intensity physical activity.

All models adjusted for sex, highest parental education level, number of siblings, number of parents and study site. Adjusted *R*-squared for all models = 0.11.

The regression estimates from the traditional regression analysis were inconsistent across the models. This demonstrates that the choice of omitted behaviour may have substantial influence on the interpretation of the relationships between the remaining behaviours and zBMI. Moreover, the regression coefficients for sleep and sedentary time varied from positive to negative, depending on the model. Figures 2 - 5 depict the inconsistency of the regression coefficients across traditional regression models and how they differ from the regression coefficients obtained from compositional models. The inconsistent findings from traditional regression models demonstrate that results from traditional analyses are unreliable when raw untransformed minutes are used as activity behaviour inputs.

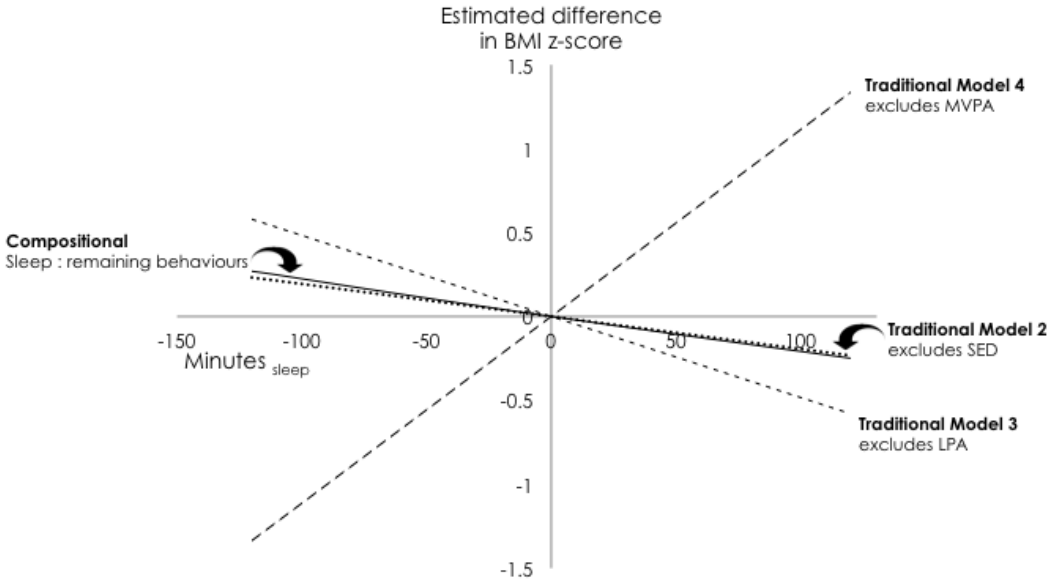
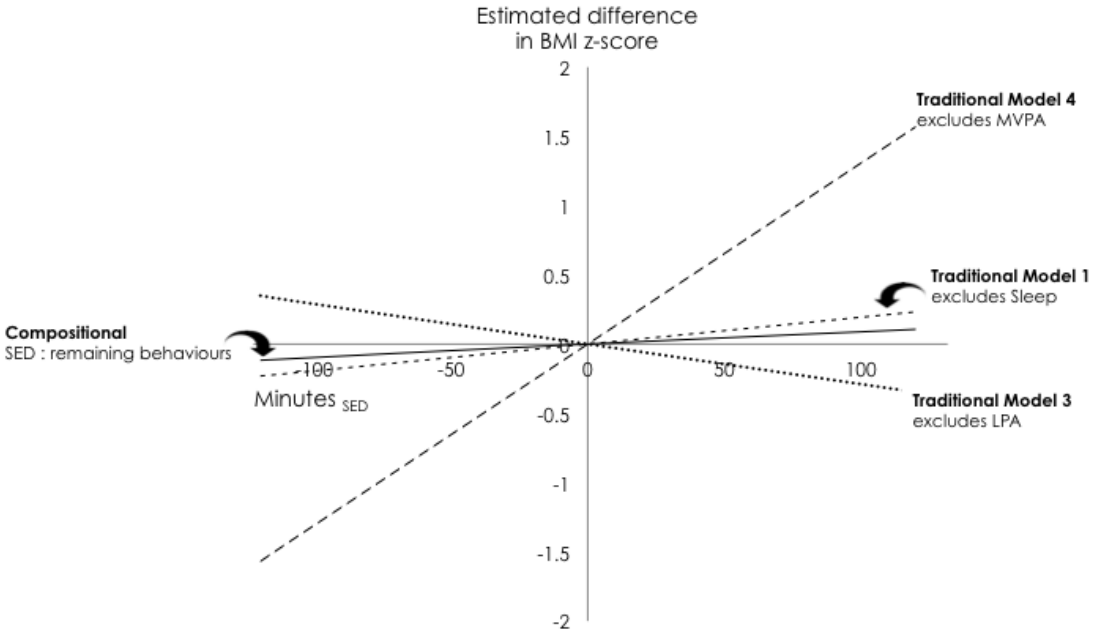


Figure 2: The Relationship Between Sleep and zBMI: Comparison between Compositional and Traditional Regression Models.

SED: Sedentary Time; LPA: Light-Intensity Physical Activity; MVPA: Moderate-to-Vigorous-Intensity Physical Activity.

Difference in Minutes Modelled Around the Population Mean Composition of (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45.



1

2

3 **Figure 3.** The Relationship Between Sedentary Time and zBMI: Comparison between

4 Compositional and Traditional Regression Models. SED: Sedentary Time; LPA: LPA:

5 Light-Intensity Physical Activity; MVPA: Moderate-to-Vigorous-Intensity Physical

6 Activity. Difference in Minutes Modelled Around the Population Mean Composition of

7 (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45

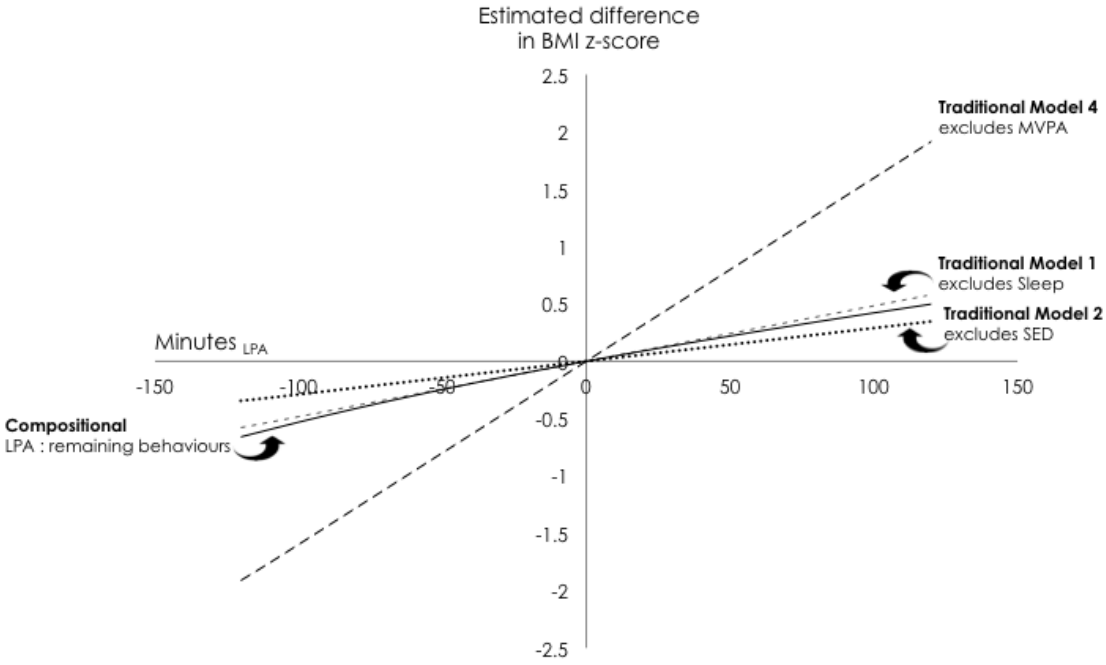


Figure 4. The Relationship Between Light Physical Activity and zBMI: Comparison between Compositional and Traditional Regression Models. SED: Sedentary Time; LPA: Light-Intensity Physical Activity; MVPA: Moderate-to-Vigorous-Intensity Physical Activity. Difference in Minutes Modelled Around the Population Mean Composition of (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45.

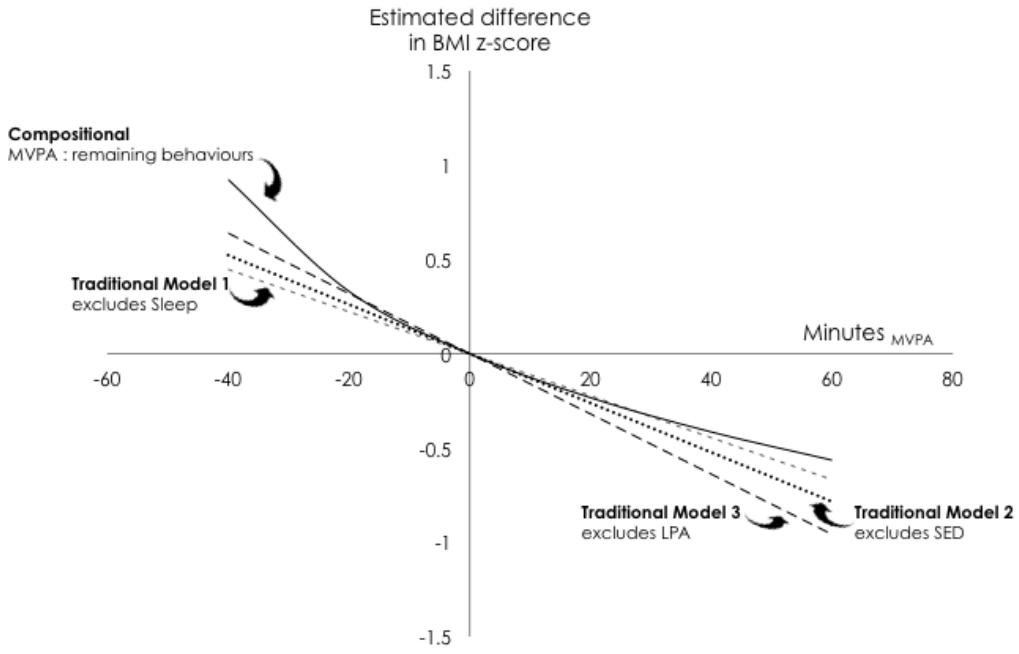


Figure 5. The Relationship Between MVPA and zBMI: Comparison between Compositional and Traditional Regression Models. SED: Sedentary Time; LPA: Light-Intensity Physical Activity; MVPA: Moderate-to-Vigorous-Intensity Physical Activity. Difference in Minutes Modelled Around the Population Mean Composition of (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45

5 Comments

While the statistical issue presented by the singularity of daily activity data has been acknowledged in previous literature,[31-33] there has been little consensus on how this might be addressed. In fact, previous research has overwhelmingly overlooked the compositional nature of time use data, and has considered daily behaviours as individual, absolute quantities. Traditional regression does not account for the closed nature of time use and the ensuing co-dependence of daily behaviours, and, consequently, findings may be spurious.[8, 13] The incontestable inability to account for all activity behaviours due to their perfect multi-collinearity is the main limitation of the traditional (non-compositional) regression models. Omitting one or more behaviours to be able to run a traditional regression analysis has been a widely used approach that ignores the true compositional nature of the data and seems to result in inconsistent regression estimates. The potential for erroneous results from traditional regression was demonstrated in this study, with estimates for sleep and sedentary time indicating a positive relationship with zBMI in some models and an inverse relationship with zBMI in other models. Such inconsistent results have the potential to undermine the credibility of academic and public health messages. For example, findings from traditional Models 1 and 4 indicate public policy should focus on the reduction of sedentary time in children, whilst findings from Model 3 indicate higher sedentary time should be encouraged to potentially reduce adiposity. Model checks (VIF) detected multi-collinearity issues for Model 4, however, VIF values for the remaining models were

below acceptable thresholds and were therefore unable to explain the inconsistencies between Model 1 and 3. This suggests that VIF is not an acceptable diagnostic indicator for compositional data. Uncertainty regarding the role of sedentary time is commonly observed in contemporary health research, with some studies finding strong associations with adiposity,[34-37] and other studies finding none.[38-41] Contemporary confusion regarding the relationship between activity behaviours and health may be a consequence of a flawed approach to statistical analysis.

Findings from the compositional data analysis (based on *ilr* coordinates) applied in this study correspond to results obtained from some traditional regression models used in this study and also some evidence from previous research.[42-45] However, because traditional models are unable to adjust for all remaining daily behaviours, outcomes may be unreliable, and are not directly comparable to outcomes from the compositional models. Furthermore, unlike compositional regression, traditional regression models are unable to detect asymmetry in associations depending on whether a behaviour is increased or decreased, or whether associations differ at various daily time-use compositions. Traditional regression can therefore not discern between the importance of maintaining or increasing a behaviour, or whether change in behaviour has a stronger association with adiposity for children with differing time-use compositions (e.g. active

children compared with sedentary children). Such considerations are important for informing public health messages and potential intervention strategies.

This study's intention was primarily to demonstrate compositional data analysis of daily activity data and interpret the findings in a meaningful manner. Therefore, analyses were carried out on the complete international dataset, however, it must be remembered that the data used were cross-sectional, and therefore causality cannot be inferred. To further investigate the relationship between activity and health, and to guide the planning of interventions and public health policy, future compositional data analyses should be performed on longitudinal data, and may be stratified by sex and be country-specific. In addition, compositional data analysis techniques for other statistical applications should be explored (e.g., isotemporal substitution, cluster analysis, principal component analysis), and include non-compositional lifestyle behaviours, such as diet quality.

The inadequacy of traditional linear regression models for data of a closed, and therefore relative nature, is well acknowledged. However, in health research, the closed nature of daily time use has been largely ignored. This study demonstrates the potential for incorrect outcomes from models commonly applied in the field. The findings imply that previous studies, including evidence accepted as being high level such as

systematic reviews, which agglomerate such studies, may be erroneous and cannot be trusted. The implications of the application of compositional data analysis to health research are quite profound. Since almost all previous analyses of the associations between time use and health outcomes have used methods incompatible with compositional data, they are all to some extent vitiated, and should be interpreted with caution. A shift towards an integrated, compositional data analysis approach, where all daily activity behaviours are considered, should be a priority. Not until robust research methodologies are implemented can valid estimates of the health associations of daily activity behaviours be made, and the mortality and economic impact of sleep, sedentary time, and physical inactivity be assessed.

Funding acknowledgements and declaration of conflicting interests

This work was supported by an Australian Government Research Training Program Scholarship [DD], the National Heart Foundation (100188) [CM] and partially supported by the Spanish Ministry of Economy and Competitiveness under the project CODA-RETOS (MTM2015-65016- C2-1(2)-R) [JAMF]. The example dataset used in this study is from the International Study of Childhood Obesity, Lifestyle and Environment (ISCOLE), which was funded by The Coca-Cola Company. The funders had no role in the design and conduct of the study, data collection, management, analysis and interpretation of the data; and decision to publish, preparation, review or

approval of this manuscript. DD, TES, JAMF and ZP are not part of ISCOLE group. The authors declare that they have no competing interests. The authors gratefully acknowledge the constructive comments of the referees which have undoubtedly helped to improve the quality of the paper.

Data availability statement

The data that support the findings of this study are available from Peter T. Katzmarzyk (Peter.Katzmarzyk@pbrc.edu) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Pennington Biomedical Research Center.

References

1. Lee I-M, Shiroma EJ, Lobelo F, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*. 2012; 380: 219-29.

2. Ding D, Lawson KD, Kolbe-Alexander TL, et al. The economic burden of physical inactivity: a global analysis of major non-communicable diseases. *The Lancet*. 2016; 388: 1311-24.

3. Hirshkowitz M, Whiton K, Albert SM, et al. National Sleep Foundation's sleep time duration recommendations: methodology and results summary. *Sleep Health*. 2015; 1: 40-3.

4. Wilmot EG, Edwardson CL, Achana FA, et al. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. *Diabetologia*. 2012; 55: 2895-905.

5. Chaput J-P, Gray CE, Poitras VJ, et al. Systematic review of the relationships between sleep duration and health indicators in school-aged children and youth 1. *Appl Physiol Nutr Metab*. 2016; 41: S266-S82.

6. Carson V, Hunter S, Kuzik N, et al. Systematic review of sedentary behaviour and health indicators in school-aged children and youth: an update 1. *Appl Physiol Nutr Metab*. 2016; 41: S240-S65.

7. Rezende LFM, Sá TH, Mielke GI, et al. All-Cause Mortality Attributable to Sitting Time: Analysis of 54 Countries Worldwide. *Am J Prev Med*. 2016; 51: 253-63.
8. Pedišić Ž. Measurement issues and poor adjustments for physical activity and sleep undermine sedentary behaviour research—the focus should shift to the balance between sleep, sedentary behaviour, standing and activity. *Kinesiology*. 2014; 46: 135-46.
9. Aitchison J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B (Methodological)*. 1982; 44: 139-77.
10. Barceló-Vidal C and Martín-Fernández J-A. The mathematics of compositional analysis. *Austr J Stat*. 2016; 45: 57-71.
11. Pawlowsky-Glahn V, Egozcue JJ and Tolosana-Delgado R. *Modeling and analysis of compositional data*. John Wiley & Sons, 2015.
12. Martín-Fernández J-A, Daunis-i-Estadella J and Mateu-Figueras G. On the interpretation of differences between groups for compositional data. *SORT-Statistics and Operations Research Transactions*. 2015; 39: 231-52.
13. Pearson K. 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*. 1896; 60: 489-98.
14. Mateu-Figueras G, Pawlowsky-Glahn V and Egozcue J. The Principle of Working on Coordinates. In: Pawlowsky-Glahn V and Buccianti A, (eds.).

Compositional Data Analysis: Theory and Applications. Chichester, UK: John Wiley & Sons, 2011, p. 29-42.

15. Palarea-Albaladejo J and Martín-Fernández JA. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometr Intell Lab*. 2015; 143: 85-96.

16. Egozcue JJ and Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. *Math Geol*. 2005; 37: 795-828.

17. Hron K, Filzmoser P and Thompson K. Linear regression with compositional explanatory variables. *J Appl Stat*. 2012; 39: 1115-28.

18. Chastin SF, Palarea-Albaladejo J, Dontje ML, et al. 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*. 2015; 10: e0139984.

19. Leite MLC. Applying compositional data methodology to nutritional epidemiology. *Stat Methods Med Res*. 2016; 25: 3057-65.

20. Mert MC, Filzmoser P, Endel G, et al. Compositional data analysis in epidemiology. *Stat Methods Med Res*. Epub ahead of print 6 October 2016: DOI: 10.1177/0962280216671536.

21. Tsilimigras MC and Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol*. 2016; 26: 330-5.

22. Carson V, Tremblay MS, Chaput J-P, et al. Associations between sleep duration, sedentary time, physical activity, and health indicators among Canadian children and youth using compositional analyses 1. *Appl Physiol Nutr Metab*. 2016; 41: S294-S302.
23. Muller I, Hron K, Fiserova E, et al. Interpretation of Compositional Regression with Application to Time Budget Analysis. *arXiv preprint*. 2016: arXiv:1609.07887.
24. Katzmarzyk PT, Barreira TV, Broyles ST, et al. The international study of childhood obesity, lifestyle and the environment (ISCOLE): Design and methods. *BMC Public Health*. 2013; 13: E900.
25. Tudor-Locke C, Barreira TV, Schuna JM, et al. Improving wear time compliance with a 24-hour waist-worn accelerometer protocol in the International Study of Childhood Obesity, Lifestyle and the Environment (ISCOLE). *Int J Behav Nutr Phys Act*. 2015; 12: 1-9.
26. Tudor-Locke C, Barreira TV, Schuna Jr JM, et al. Fully automated waist-worn accelerometer algorithm for detecting children's sleep-period time separate from 24-h physical activity or sedentary behaviors. *Appl Physiol Nutr Metab*. 2013; 39: 53-7.
27. Barreira TV, Schuna Jr JM, Mire EF, et al. Identifying children's nocturnal sleep using 24-h waist accelerometry. *Med Sci Sports Exerc*. 2015; 47: 937-43.
28. Evenson KR, Catellier DJ, Gill K, et al. Calibration of two objective measures of physical activity for children. *J Sports Sci*. 2008; 26: 1557-65.

29. de Onis M, Onyango A, Borghi E, et al. Development of a WHO growth reference for school-aged children and adolescents. *Bull World Health Organ*. 2007.
30. Fox J and Monette G. Generalized collinearity diagnostics. *J Am Stat Assoc*. 1992; 87: 178-83.
31. Augustin NH, Mattocks C, Cooper AR, et al. Modelling fat mass as a function of weekly physical activity profiles measured by Actigraph accelerometers. *Physiol Meas*. 2012; 33: 1831.
32. Hunt E, McKay E, Fitzgerald A, et al. Time use and daily activities of late adolescents in contemporary Ireland. *J Occup Sci*. 2014; 21: 42-64.
33. Maher C, Olds T, Mire E, et al. Reconsidering the sedentary behaviour paradigm. *PLoS ONE*. 2014; 9: e86403.
34. Hussey J, Bell C, Bennett K, et al. Relationship between the intensity of physical activity, inactivity, cardiorespiratory fitness and body composition in 7–10-year-old Dublin children. *Br J Sports Med*. 2007; 41: 311-6.
35. Kennedy K, Shepherd S, Williams JE, et al. Activity, body composition and bone health in children. *Arch Dis Child*. 2013; 98: 204-7.
36. Machado-Rodrigues AM, Coelho-e-Silva MJ, Mota J, et al. Cardiorespiratory fitness, weight status and objectively measured sedentary behaviour and physical activity in rural and urban Portuguese adolescents. *J Child Health Care*. 2012; 16: 166-77.

37. Marques A, Minderico C, Martins S, et al. Cross-sectional and prospective associations between moderate to vigorous physical activity and sedentary time with adiposity in children. *Int J Obes*. 2016; 40: 28-33.

38. Loprinzi PD, Cardinal BJ, Lee H, et al. Markers of adiposity among children and adolescents: implications of the isotemporal substitution paradigm with sedentary behavior and physical activity patterns. *J Diabetes Metab Disord*. 2015; 14: 1.

39. Dowd KP, Harrington DM, Hannigan A, et al. Light-intensity physical activity is associated with adiposity in adolescent females. *Med Sci Sports Exerc*. 2014; 46: 2295-300.

40. Basterfield L, Pearce MS, Adamson AJ, et al. Physical activity, sedentary behavior, and adiposity in English children. *Am J Prev Med*. 2012; 42: 445-51.

41. Hjorth MF, Chaput J-P, Damsgaard CT, et al. Low physical activity level and short sleep duration are associated with an increased cardio-metabolic risk profile: a longitudinal study in 8-11 year old Danish children. *PLoS ONE*. 2014; 9: e104677.

42. Saunders TJ, Gray CE, Poitras VJ, et al. Combinations of physical activity, sedentary behaviour and sleep: relationships with health indicators in school-aged children and youth 1. *Appl Physiol Nutr Metab*. 2016; 41: S283-S93.

43. Tremblay MS, LeBlanc AG, Kho ME, et al. Systematic review of sedentary behaviour and health indicators in school-aged children and youth. *Int J Behav Nutr Phys Act*. 2011; 8: 98.

44. Janssen I and LeBlanc AG. Review Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. *Int J Behav Nutr Phys Act*. 2010; 7: 1-16.

45. Chen X, Beydoun MA and Wang Y. Is sleep duration associated with childhood obesity? A systematic review and meta‐analysis. *Obesity*. 2008; 16: 265-74.

Supplementary file 1

The model

As outlined in the main paper, the *ilr* model for n compositional observations $(\boldsymbol{x_i}, y_i), i = 1,2, \dots, n$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{iD})^T$ with $\sum_{j=1}^{D} x_{ij} = 1$, is

$$y_i = \beta_0 + z_i^T \beta + \epsilon_i$$

$$= \beta_0 + \sum_{j=1}^{D-1} \beta_j z_{ij} + \epsilon_i, \qquad (1)$$

where

$$z_{ij} = \sqrt{\frac{D-j}{D-j+1}} \ln\left(\frac{x_{ij}}{\left(\prod_{k=j+1}^{D} x_{ik}\right)^{1/(D-j)}}\right) \text{ for } j = 1, 2, \dots, D-1.$$

Note that covariates can be included without transform (subject to the standard multiple linear regression assumptions) into the *ilr* multiple linear regression model without alterations to the derived formula in Equation (8) of the main paper. For completeness, should there be E covariates for each observation i

$(w_{i1}, w_{i2}, \dots, w_{iE})$ to be added to the *ilr* multiple linear regression model, the model would be similarly specified as follows,

$$y_i = \beta_0 + \sum_{j=1}^{D-1} \beta_j z_{ij} + \sum_{j=1}^{E} \beta_{D-1+j} w_{ij} + \epsilon_i.$$

Relative changes in the components of x_i

Consider a relative increase in $x_{i1}$by a factor of $1 + r$ with $-1 < r < \frac{1-x_1}{x_{i1}}$. Note that r cannot take a value of $\frac{1-x_1}{x_{i1}}$ (or greater) as

$$(1 + r)x_{i1} < 1$$

$$\Rightarrow r x_{i1} < 1 - x_{i1}$$

$$\Rightarrow r < \frac{1-x_1}{x_{i1}}$$

To maintain the compositional components' sum to unity when a relative increase in $x_{i1}$by a factor of $1 + r$ is applied, the remaining components can be reduced using a factor of $1 - s$. We derive the value of s below.

The individual components are adjusted as per below:

$$x_{i1} \rightarrow (1+r)x_{i1} = x_{i1}^*$$

$$x_{i2} \rightarrow (1+r)x_{i2} = x_{i2}^*$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$x_{iD} \rightarrow (1+r)x_{iD} = x_{iD}^*$$

Note that $\sum_{j=1}^{D} x_{ij} = 1$ and $\sum_{j=1}^{D} x_{ij}^* = 1$. Also note therefore $\sum_{j=2}^{D} x_{ij} = 1 - x_{i1}$.

Now to find an expression for *s*, consider the following:

$$\sum_{j=1}^{D} x_{ij}^* = 1$$

$$\Rightarrow (1+r)x_{i1} + \sum_{j=2}^{D}(1-s)x_{ij} = 1$$

$$\Rightarrow (1+r)x_{i1} + (1-s)\sum_{j=2}^{D} x_{ij} = 1$$

$$\Rightarrow (1+r)x_{i1} + (1-s)(1-x_{i1}) = 1$$

$$\Rightarrow r x_{i1} = s(1-x_{i1})$$

$$\Rightarrow s = r\frac{x_{i1}}{1-x_{i1}}$$

Estimated change in the outcome for an increase in the first compositional part and equal relative reductions in the remaining components

Consider a new observation

$$\boldsymbol{x}_0 = (x_{01}, x_{02}, \ldots, x_{0D})^T$$

and the corresponding *ilr* coordinates

$$\boldsymbol{z}_0 = (z_{01}, z_{02}, \ldots, z_{0(D-1)})^T.$$

Now let us consider a new set of predictor variables

$$\boldsymbol{x}_0^* = ((1+r)x_{01}, (1-s)x_{02}, \ldots, (1-s)x_{0D})^T$$

and the corresponding *ilr* coordinates

$$\boldsymbol{z}_0^* = (z_{01}^*, z_{02}^*, \ldots, z_{0(D-1)}^*)^T$$

where

$$z_{01}^* = \sqrt{\frac{D-1}{D}} \ln\left(\frac{x_{01}^*}{(\prod_{k=2}^{D} x_{0k}^*)^{1/(D-1)}}\right)$$

$$= \sqrt{\frac{D-1}{D}} \ln\left(\frac{(1+r)x_{01}}{\left((1-s)^{D-1}(\prod_{k=2}^{D} x_{0k})\right)^{1/(D-1)}}\right)$$

$$= \sqrt{\frac{D-1}{D}} \ln\left(\frac{(1+r)}{(1-s)}\frac{x_{01}}{(\prod_{k=2}^{D} x_{0k})^{1/(D-1)}}\right), and$$

$$z_{0j}^{*} = \sqrt{\frac{D-j}{D-j+1}}\,\ln\left(\frac{x_{0j}^{*}}{\left(\prod_{k=j+1}^{D} x_{0k}^{*}\right)^{1/(D-1)}}\right) \; for\; j = 2,3,\ldots,D-1$$

$$= \sqrt{\frac{D-j}{D-j+1}}\,\ln\left(\frac{(1-s)x_{0j}}{\left((1-s)^{D-j}\left(\prod_{k=j+1}^{D} x_{0k}\right)\right)^{1/(D-j)}}\right)$$

$$= \sqrt{\frac{D-j}{D-j+1}}\,\ln\left(\frac{(1-s)}{(1-s)}\frac{x_{0j}}{\left(\prod_{k=j+1}^{D} x_{0k}\right)^{1/(D-j)}}\right)$$

$$= \sqrt{\frac{D-j}{D-j+1}}\,\ln\left(\frac{x_{0j}}{\left(\prod_{k=j+1}^{D} x_{0k}\right)^{1/(D-j)}}\right)$$

$= z_{01}$.

That is, there is no change in the *ilr* coordinates for $j = 2,3,\ldots,D-1$ and

$$\boldsymbol{z}_0^{*} = (z_{01}^{*}, z_{02}^{*}, \ldots, z_{0(D-1)}^{*})^{T}.$$

The estimated outcome for predictors $\boldsymbol{x}_0$ is

$$\hat{y}_0 = \hat{\beta}_0 + z_0^T \hat{\beta}$$

and the estimated outcome for predictors $\boldsymbol{x_0^*}$ is

$$\hat{y}_0^* = \hat{\beta}_0 + z_0^{*T} \hat{\beta}.$$

Therefore the estimated change in the predicted outcome going from predictors $\boldsymbol{x}_0$ to $\boldsymbol{x_0^*}$ is

$$\Delta\hat{y} = \hat{y}_0^* - \hat{y}_0$$

$$= \hat{\beta}_0 + z_0^{*T}\hat{\beta} - \hat{\beta}_0 + z_0^T\hat{\beta}$$

$$= \left(\hat{\beta}_1 z_{01}^* + \sum_{j=2}^{D-1} \hat{\beta}_j z_{0j}\right) - \left(\hat{\beta}_1 z_{01} + \sum_{j=2}^{D-1} \hat{\beta}_j z_{0j}\right)$$

$$= \hat{\beta}_1 (z_{01}^* - z_{01})$$

$$= \hat{\beta}_1 \left(\sqrt{\frac{D-1}{D}} \ln\left(\frac{(1+r)}{(1-s)} \frac{x_{01}}{\left(\prod_{k=2}^{D} x_{0k}\right)^{1/(D-1)}}\right) - \sqrt{\frac{D-1}{D}} \ln\left(\frac{x_{01}}{\left(\prod_{k=2}^{D} x_{0k}\right)^{1/(D-1)}}\right)\right)$$

$$= \hat{\beta}_1 \sqrt{\frac{D-1}{D}} \ln \left(\frac{(1+r)}{(1-s)} \frac{x_{01}}{(\prod_{k=2}^{D} x_{0k})^{1/(D-1)}} \Big/ \frac{x_{01}}{(\prod_{k=2}^{D} x_{0k})^{1/(D-1)}} \right)$$

$$= \hat{\beta}_1 \sqrt{\frac{D-1}{D}} \ln \left(\frac{1+r}{1-s} \right)$$ as shown in equation (8) of the main paper.

Confidence interval for $\Delta\hat{y}$

The standard error of $\Delta\hat{y}$ *is*

$$\mathrm{SE}(\Delta\hat{y}) = \sigma \sqrt{\frac{D-1}{D}} ln \left(\frac{1+r}{1-s} \right) \sqrt{(Z^T Z)^{-1}_{(1)}}$$

as

$$\mathrm{var}(\Delta\hat{y}) = \mathrm{var} \left(\hat{\beta}_1 \sqrt{\frac{D-1}{D}} \ln \left(\frac{1+r}{1-s} \right) \right)$$

$$= \left(\sqrt{\frac{D-1}{D}}\ln\left(\frac{1+r}{1-s}\right)\right)^2 \text{var}\left(\hat{\beta}_1\right)$$

$$= \sigma^2\left(\sqrt{\frac{D-1}{D}}ln\left(\frac{1+r}{1-s}\right)\right)^2 (Z^TZ)^{-1}_{(1)}$$

where Z is the design matrix of the model in (1) and $\text{A}_{(1)}$ denotes the first diagonal element of A.

Therefore the $100(1-\alpha)\%$ confidence interval for Δy is

$$\left(\Delta\hat{y} - t_{\alpha/2,n-D}\sigma\sqrt{\frac{D-1}{D}}ln\left(\frac{1+r}{1-s}\right)\sqrt{(Z^TZ)^{-1}_{(1)}}, \Delta\hat{y} + t_{\alpha/2,n-D}\sigma\sqrt{\frac{D-1}{D}}ln\left(\frac{1+r}{1-s}\right)\sqrt{(Z^TZ)^{-1}_{(1)}}\right).$$

Supplementary file 2

Daily activity composition expressed as isometric log-ratio coordinates

A four-part daily activity composition consisting of: sleep (*sleep*), sedentary time (*SED*), light-intensity physical activity (*LPA*) and moderate-to-vigorous-intensity physical activity (*MVPA*), can be expressed by a set of three isometric log-ratio coordinates [z_{i1}, z_{i2}, z_{i3}] as follows:

$$z_{i1} = \sqrt{\frac{3}{4}} \ln\left(\frac{sleep_i}{\sqrt[3]{SED_i \cdot LPA_i \cdot MVPA_i}}\right),$$

$$z_{i2} = \sqrt{\frac{2}{3}} \ln\left(\frac{SED_i}{\sqrt{LPA_i \cdot MVPA_i}}\right) \text{ and}$$

$$z_{i3} = \sqrt{\frac{1}{2}} \ln\left(\frac{LPA_i}{MVPA_i}\right).$$

To estimate an outcome (e.g., body mass index z-score [zBMI]), a multiple linear regression model can be constructed with the above log-ratio coordinates as the explanatory variables:

$$zBMI_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + covariates_i + \epsilon_i$$

The coefficient β_1 corresponds to z_{i1}, which is the log-ratio of $sleep_i$, to the geometric mean of the remaining behaviours (*SED*$_i$, *LPA*$_i$ and *MVPA*$_i$).

Permutation of the compositional parts iteratively to place each behaviour as the first part of the composition, and then applying the above isometric log-ratio coordinates allows four linear models to be created, with each model's z_{i1} representing the first (permuted) behaviour in relation to the geometric mean of the remaining behaviours.

$$Composition\ 1 = [Sleep, SED, LPA, MVPA]$$

$$Composition\ 2 = [SED, LPA, MVPA, sleep]$$

$$Composition\ 3 = [LPA, MVPA, sleep, SED]$$

$$Composition\ 4 = [MVPA, sleep, SED, LPA\ ]$$

Predicting change in zBMI using the linear model

The predictive model using Composition 1 becomes:

$$\widehat{zBMI}_i = \hat{\beta}_0 + \ \hat{\beta}_1 z_{i1} + \hat{\beta}_2 z_{i2} + \hat{\beta}_3\, z_{i3} + covariates_i.$$

We define the *ilr* coordinates [z_{i1}, z_{i2}, z_{i3}] as described above.

As compositional data are relative, the predictions from the model must be made relative to a starting/reference activity behaviour composition. In our example, we select the population mean activity behaviour composition as the starting point, expressed as proportions with a closure constant of 1.

To predict zBMI for a new composition where our first compositional part (*sleep*) has changed, we multiply $sleep_{mean}$ by a constant (1 +/- r) (e.g., to increase $sleep_{mean}$ by 5%, r = 0.05, and we multiply $sleep_{mean}$ by 1+ r = 1.05). However, due to the constant sum constraint of daily activity data, the remaining behaviours must be changed accordingly. The remaining compositional parts are therefore all simultaneously multiplied by another constant (1 -/+ s), specifically derived to maintain the total sum of all parts to 1 (see Supplementary file 1 for details). By multiplying all the remaining parts by the same constant (in our example the remaining parts are decreased, therefore each remaining part is multiplied by 1 - s), the remaining log ratio coordinates (z_{i2}, z_{i3}) are kept constant (as both numerator and denominator of the log ratio coordinates are multiplied by the same amount [1 - s]).

Therefore, we can use the linear model to predict zBMI for a change in the daily activity composition by the constant k,

$$\text{where } k = \frac{1+r}{1-s}$$

and r and s are defined as described in Supplementary file 1.

The predictive model for our example becomes (N.B.: the following is a worked example of the proofs given in Supplementary file 1):

$$\widehat{zBMI}(k) = \hat{\beta}_0 + \hat{\beta}_1\sqrt{\frac{3}{4}}\ln\left(\frac{sleep}{\sqrt[3]{SED\cdot LPA\cdot MVPA}}\cdot k\right) + \hat{\beta}_2\sqrt{\frac{2}{3}}\ln\left(\frac{SED}{\sqrt{LPA\cdot MVPA}}\right) + \hat{\beta}_3\sqrt{\frac{1}{2}}\ln\left(\frac{LPA}{MVPA}\right)$$

$$= \hat{\beta}_0 + \hat{\beta}_1\sqrt{\frac{3}{4}}\ln\left(\frac{sleep}{\sqrt[3]{SED\cdot LPA\cdot MVPA}}\right) + \hat{\beta}_1\sqrt{\frac{3}{4}}\ln k + \hat{\beta}_2\sqrt{\frac{2}{3}}\ln\left(\frac{SED}{\sqrt{LPA\cdot MVPA}}\right) + \hat{\beta}_3\sqrt{\frac{1}{2}}\ln\left(\frac{LPA}{MVPA}\right)$$

$(expanding\ out\ the\ first\ ratio\ using\ \log law : \ln(ab) = \ln(a) + \ln(b))$

$$= \hat{\beta}_0 + \hat{\beta}_1\sqrt{\frac{3}{4}}\ln\left(\frac{sleep}{\sqrt[3]{SED\cdot LPA\cdot MVPA}}\right) + \hat{\beta}_2\sqrt{\frac{2}{3}}\ln\left(\frac{SED}{\sqrt{LPA\cdot MVPA}}\right) + \hat{\beta}_3\sqrt{\frac{1}{2}}\ln\left(\frac{LPA}{MVPA}\right) + \hat{\beta}_1\sqrt{\frac{3}{4}}\ln k$$

$(rearrange\ the\ order\ so\ that\ it's\ easy\ to\ see\ the\ first\ bit\ can\ be\ replaced\ by\ 'zBMI(k)')$

$$= \widehat{zBMI} + \hat{\beta}_1\sqrt{\frac{3}{4}}\ln k$$

Therefore, the predicted change in zBMI is $\hat{\beta}_1\sqrt{\frac{3}{4}}\ln k$.

Here $k = \frac{1+r}{1-s}$, i.e., with an increase in the first behaviour (sleep), *sleep* is multiplied by 1 + *r*, and the remaining behaviours (*SED*, *LPA*, *MVPA*), represented by their geometric mean, are each multiplied by 1 - *s*.

As stated earlier, by multiplying each of the remaining behaviours simultaneously by 1 - *s*, it is assumed that the log-ratio coefficients z_{i2} and z_{i3} are held constant.

The value of 1 - *s* can be derived from 1 + *r*, when we consider daily activity data to be constrained to $C = 1$, i.e., the daily activity components are expressed as proportions.

The reference/starting daily composition can be expressed as:

$$sleep + remaining = 1.$$

However, we would like to predict an outcome (zBMI) for a new composition, where sleep is increased by a factor of 1+*r*, and remaining behaviours are decreased by a factor of 1-*s*. The new composition can be expressed as:

$$sleep\ (1 + r) + remaining\ (1 - s) = 1;$$

$$\therefore sleep + r \cdot sleep + remaining - remaining \cdot s = 1$$

$$\therefore sleep + remaining + r \cdot sleep - remaining \cdot s = 1.$$

As sleep + remaining = 1, therefore:

$$(1 + r) \cdot sleep - remaining \cdot s = 1;$$

$$\therefore r \cdot sleep - remaining \cdot s = 0$$

$$\therefore remaining \cdot s = r \cdot sleep;$$

$$\therefore s = r \cdot \frac{sleep}{remaining}.$$

Application to the data presented in the manuscript:

The mean daily activity composition (described by geometric means, closed to 1440 minutes) was:

sleep = 539; *SED* = 525; *LPA* = 320; *MVPA* = 57.

This can be expressed as a set of proportions [0.374, 0.364, 0.222, 0.040], which are closed to 1.

Expressed as a proportion, $sleep_{mean}$ = 0.374, therefore the *remaining* components (expressed as proportions) must together equal: 1 (*the total day*) - 0.374 ($sleep_{mean}$).

If we are interested in change in zBMI when sleep is increased (relatively) by 5% (r = 0.05), we can calculate s, using the formula above. Specifically, $s = 0.05 \cdot \frac{0.374}{1-0.374} = 0.03$.

We now have our constant r (=0.05) which is the relative increase in $sleep_{mean}$, and our constant s (=0.03), which is the relative decrease in each of the remaining behaviours. Using r and s, we can create k, the constant which is applied to the first log ratio coordinate (z_{i1}). In our example, $k = \frac{1+r}{1-s} = \frac{1.05}{0.97}$.

Earlier we showed that predicted change in zBMI was equal to:

$$\triangle \widehat{zBMI} = \hat{\beta}_1 \cdot \sqrt{\frac{3}{4}} \cdot \ln k, \text{ where } k = \frac{1+r}{1-s}.$$

Now, we substitute $k = \frac{1.05}{0.97}$ into our "change" equation, and use the value for $\hat{\beta}_1 = -0.82$ from the ilr_{sleep} linear regression model fit (see Table 1 in main paper).

The change in $\widehat{zBMI}$ is

$$\hat{\beta}_1 \sqrt{\frac{3}{4}} \ln \frac{1.05}{0.97} = -0.82 \cdot \sqrt{\frac{3}{4}} \cdot \ln \frac{1.05}{0.97} = -0.056.$$

Therefore, predicted zBMI has decreased by 0.056 units when sleep is increased from the reference/starting composition (in this case, the population-mean composition) by 1.05 or 5%, relative to remaining behaviours.

The change in daily activity composition can be interpreted in minutes. Mean *sleep* = 539 minutes, therefore a 5% relative decrease in sleep is a decrease of 27 minutes.

Conversely, if sleep is decreased in relative terms by 5% (r = - 0.05), then

$$k = \frac{1-0.05}{1+0.03}.$$

Therefore,

$$\text{change in } \widehat{zBMI} = -0.82 \cdot \sqrt{\frac{3}{4}} \cdot \ln\frac{1-0.05}{1+0.03} = 0.057.$$

zBMI is predicted to increase by 0.057 when sleep is decreased by 5% from the reference/starting composition, relative to remaining behaviours.

We can use the linear model to predict how much the first compositional part must change for a specific change in a continuous outcome (e.g., using the composition above, we can predict the change in $sleep_{mean}$ associated with a decrease in zBMI of 0.1 units). To use the model for this prediction, we must first isolate *r*.

We have established that:

$$\Delta\,\hat{y}\,(e.g.zBMI) = \hat{\beta}_1\sqrt{\frac{3}{4}}\ln k,$$

where (for a 4-part composition), *y* = the predicted variable (e.g., zBMI), $\hat{\beta}_1$ = the coefficient of the first log-ratio regression coefficient (one behaviour: remaining day, therefore contains all relative information regarding $\bar{x}_1$, the first compositional part of the mean composition), and $k = (1+r)/(1-s)$, where $s = r(\bar{x}_1 x/1-\bar{x}_1 x)$.

Therefore, rearranging this formula, we can isolate *k*

$$\ln k = \frac{\Delta\hat{y}}{\hat{\beta}_1 \cdot \sqrt{\frac{3}{4}}}$$

$$\therefore k = e^{\frac{\Delta\hat{y}}{\hat{\beta}_1 \cdot \sqrt{\frac{3}{4}}}}.$$

To calculate *r*, we can use that $k = (1+r)/(1-s)$, where $s = r\bar{x}_1/(1-\bar{x}_1)$.

Therefore,

$$k = \frac{1+r}{1-r\left(\frac{\bar{x}_1}{1-\bar{x}_1}\right)}$$

where we can isolate *r*:

$$r = \frac{-1+k}{k\left(\frac{\bar{x}_1}{1-\bar{x}_1}\right)+1.}$$

Now we can calculate *r* by substituting in *k* from above, using the $\hat{\beta}_1$ from the regression model, and the $\Delta\hat{y}$ of interest to calculate *k*.

To express *r* as change in minutes from a reference x_1, e.g., $\bar{x}_1$, $r(minutes) = r \cdot \bar{x}_1 \cdot 1440$.