Regression for Compositions: Logit Models versus Log-ratio Transformation of Data

useR! 2024, Salzburg

David Firth
University of Warwick

Firth, D and Sammut, F (2023). Analysis of composition on the original scale of measurement.
arXiv:2312.10548

Compositional data analysis

A very active area of applied multivariate statistics.

Composition: the relative sizes of parts of a thing.

Examples:

- minerals found in samples of rock or sediment
- vote shares in multi-party elections
- household expenditure patterns
- personal time-use data
- microbiome data (e.g., human gut)

etc., etc.

Main aim here: show how to use `compos::colm()` to fit compositional logit models.

Package in development at https://github.com/DavidFirth/compos — will be ready for CRAN 'soon'!

First, though, brief background on the models and associated statistical methods.

Compositional logit model

Non-negative measurements

$$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iD}) = T_i(P_{i1}, \ldots, P_{iD}) = T_i\mathbf{P}_i \qquad (i = 1, \ldots, n),$$

with $\sum_{j=1}^{d} P_{ij} = 1$.

The measurements are always extensive variables, so focus on arithmetic mean

$$E(\mathbf{Y}_i) = E(T_i\mathbf{P}_i) = \tau_i\boldsymbol{\pi}_i$$

(e.g., Cox & Donnelly, *Principles of Applied Statistics*, ch. 4)

Compositional logit (CL) model then takes the form

$$\log(\pi_{ij}/\pi_{ik}) = \mathbf{x}_i^T(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k)$$

for all j and k in $\{1, \ldots, D\}$.

CL model:

$$\log(\pi_{ij}/\pi_{ik}) = \mathbf{x}_i^T(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k)$$

Two aspects to note:

- Only *differences* (or more generally linear *contrasts*) among parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D$ are identified.
- The covariate vector $\mathbf{x}_i$ is taken to be the same for every one of the logits $\log(\pi_{ij}/\pi_{ik})$. This might seem rather restrictive, but it can readily be relaxed, e.g., via *nested* CL models.

A standard alternative: Linear model for log-ratio transformed data

As in Aitchison (1986) *The Statistical Analysis of Compositional Data*.

Multivariate linear model for log-ratios:

$$E[\log(P_{ij}/P_{ik})] = \mathbf{x}_i^T(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k).$$

Compare with our CL model:

$$\log[E(P_{ij})/E(P_{ik})] = \mathbf{x}_i^T(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k).$$

(a generalized linear model, with logit link)

Logit versus log-ratio

Key advantages of the compositional logit (CL) model:

- A model for arithmetic means directly. Hence straightforward interpretation on the original scale of measurement.
- Can readily avoid problems with zeros, or with near-zero values in the data. Zeros present a major practical problem for use of the log-ratio transformation.

Model assumptions: Parametric or semi-parametric?

Parametric: The most standard assumptions are

- For **counts**: Multinomial logit models, $\mathbf{Y}_i | T_i \sim \text{multinomial}(T_i, \boldsymbol{\pi}_i)$. In *R*, use `glm()` or `nnet::multinom()`. (and there are others)
- For **continuous data**:
  - Beta and Dirichlet models. Two well-established CRAN packages are **betareg** and **DirichletReg**.
  - logistic normal models (Aitchison, 1986) in which the error vectors for the log-ratio linear model are multivariate normal $\text{MVN}(\mathbf{0}, \boldsymbol{\Phi})$. In *R* we can just use `lm()`.

Model assumptions: Parametric or semi-parametric?

Semi-parametric: Notice that the logistic-normal linear model represents multiplicative error:

$$Y_{ij} = \tau_i \pi_{ij} U_{ij} = \tau_i \exp(\mathbf{x}_i^T \boldsymbol{\beta}_j) U_{ij} \quad \text{with} \quad \log \mathbf{U}_i \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Phi}).$$

Relax the MVN assumption to just the first two moments: $E(\mathbf{U}_i) = \mathbf{1}, \mathrm{cov}(\mathbf{U}_i) = \boldsymbol{\Phi}$. Then (from F&S 2023):

$$E(\mathbf{P}_i) = \boldsymbol{\pi}_i; \qquad \mathrm{cov}(\mathbf{P}_i) = (\boldsymbol{\Pi}_i - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^T) \boldsymbol{\Phi} (\boldsymbol{\Pi}_i - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^T),$$

where $\boldsymbol{\Pi}_i = \mathrm{diag}(\boldsymbol{\pi}_i)$.

This generalizes a suggestion from Wedderburn (1974) for the specific case $D = 2$. Call it the generalized Wedderburn variance-covariance function.

Wedderburn (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*.

Model assumptions: Parametric or semi-parametric?

Parametric advantages:

- Full likelihood and corresponding statistical methods are available.

Semi-parametric advantages:

- Robustness to failure of distributional assumptions (only second-moment assumptions are made).
- Stability with zero or near-zero values in the data (because quasi-likelihood estimating equations are linear in the data).

The second of these is especially important in practice.

See F&S 2023 for more details, and/or the package vignette in **compos**.

compos::colm() in action: Arctic lake sediment data

Data 5. *Sand, silt, clay compositions of 39 sediment samples at different water depths in an Arctic lake*

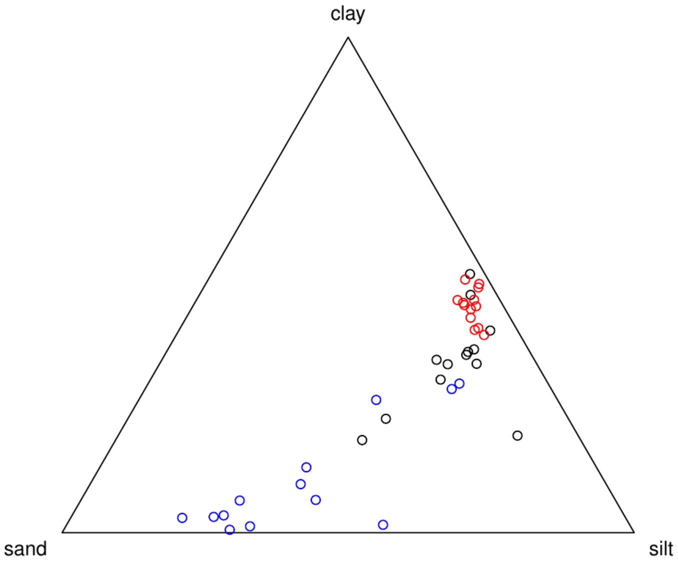
Sediment no.	*Percentages*			*Water depth (m)*	*Sediment no.*	*Percentages*			*Water depth (m)*
	Sand	*Silt*	*Clay*			*Sand*	*Silt*	*Clay*	
S1	77.5	19.5	3.0	10.4	S21	9.5	53.5	37.0	47.1
S2	71.9	24.9	3.2	11.7	S22	17.1	48.0	34.9	48.4
S3	50.7	36.1	13.2	12.8	S23	10.5	55.4	34.1	49.4
S4	52.2	40.9	6.6	13.0	S24	4.8	54.7	41.0	49.5
S5	70.0	26.5	3.5	15.7	S25	2.6	45.2	52.2	59.2
S6	66.5	32.2	1.3	16.3	S26	11.4	52.7	35.9	60.1
S7	43.1	55.3	1.6	18.0	S27	6.7	46.9	46.4	61.7
S8	53.4	36.8	9.8	18.7	S28	6.9	49.7	43.4	62.4
S9	15.5	54.4	30.1	20.7	S29	4.0	44.9	51.1	69.3
S10	31.7	41.5	26.8	22.1	S30	7.4	51.6	40.9	73.6
S11	65.7	27.8	6.5	22.4	S31	4.8	49.5	45.7	74.4
S12	70.4	29.0	0.6	24.4	S32	4.5	48.5	47.0	78.5
S13	17.4	53.6	29.0	25.8	S33	6.6	52.1	41.3	82.9
S14	10.6	69.8	19.6	32.5	S34	6.7	47.3	45.9	87.7
S15	38.2	43.1	18.7	33.6	S35	7.4	45.6	46.9	88.1
S16	10.8	52.7	36.5	36.8	S36	6.0	48.9	45.1	90.4
S17	18.4	50.7	30.9	37.8	S37	6.3	53.8	39.9	90.6
S18	4.6	47.4	48.0	36.9	S38	2.5	48.0	49.5	97.7
S19	15.6	50.4	34.0	42.2	S39	2.0	47.8	50.2	103.7
S20	31.9	45.1	23.0	47.0					

Adapted from Coakley and Rust (1968, Table 1).

This is the classic dataset from Aitchison (1986).

(Actually the table in Aitchison's book contains two errors, which seem not to have been noticed in later papers/packages. The errors are documented and corrected in **compos**.)

Arctic lake sediment data: red samples are at depth > 60.8m, blue samples are at depth < 29.4m.



Arctic lake sediments: logit model vs log-ratio model

```
library(compos); data(arctic_lake)
sediments <- arctic_lake[, c("sand", "silt", "clay")]
logdepth <- log(arctic_lake[, "depth"])

logitModel <- colm(sediments ~ logdepth)
```

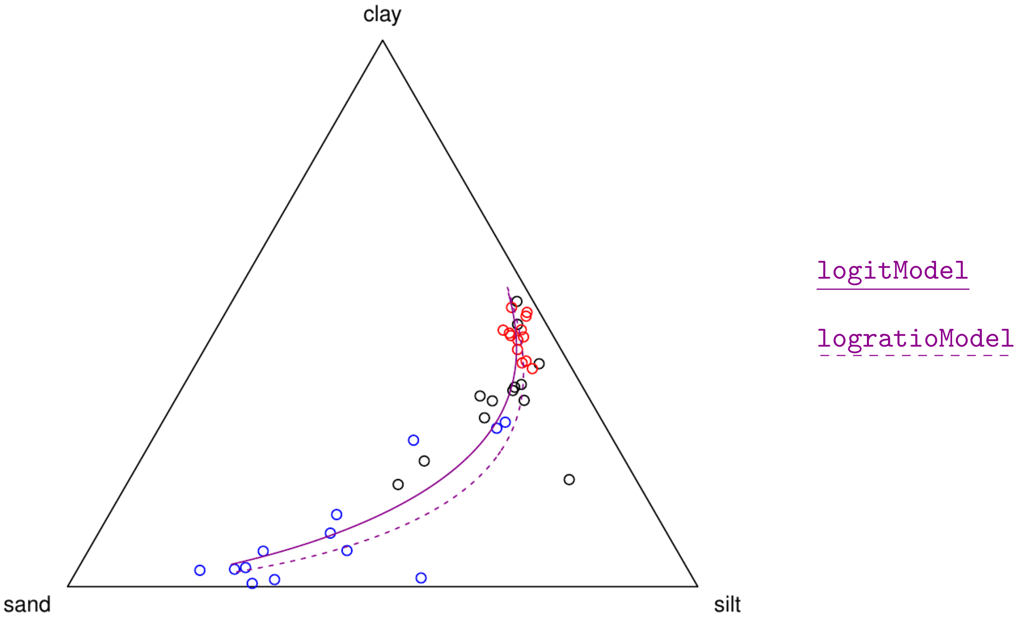
```
              sand   silt   clay
(Intercept) -0.43   0.62   0.00
logdepth    -2.48  -0.86   0.00
```

```
logratioModel <- colm(sediments ~ logdepth, method = "logratio_fit")
```

```
              sand   silt   clay
(Intercept) -0.37   0.78   0.00
logdepth    -2.74  -1.10   0.00
```

Standard errors, via `summary(logitModel)` etc., differ a bit too.

To see the influence of some *near-zero values*, we can plot the fitted values:


clay
sand
silt
logitModel
logratioModel

- Firth, D and Sammut, F (2024). Package **compos**.
  https://github.com/DavidFirth/compos
- Firth, D and Sammut, F (2023). Analysis of composition on the original scale of measurement. *arXiv:2312.10548*

‘Independence of irrelevant alternatives’

(to be included in the useR! talk *only* if there is spare time!)

In the compositional logit model, $\log(\pi_{ij}/\pi_{ik})$ does not involve other elements of $\boldsymbol{\pi}_i$.

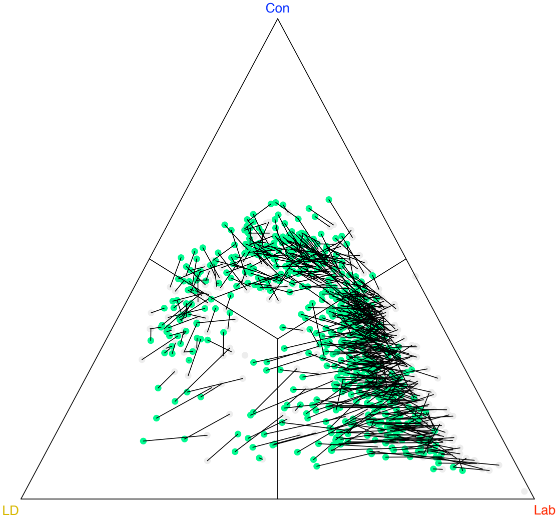
So the compositional logit model has the well known independence of irrelevant alternatives (**IIA**) property.

Depending on the application, that might be a good thing or a bad thing.

But in any case, it can always be tailored to the application via a nested sequence of compositional logit models.

This **IIA** property is related to (but also preferable to!) Aitchison’s subcompositional coherence requirement for compositional data analysis.

Nested analysis: An example



Multi-party election data.

Curtice & F (2008) use separate models for

- LD | {Con, Lab, LD}
- Lab | {Con, Lab}.