Residuals Analysis of The Dirichlet Regression Models

Rafiq H. Hijazi

Department of Statistics
United Arab Emirates University
P.O.Box 17555, Al-Ain, UAE
Email: rhijazi@uaeu.ac.ae

November 2, 2008

Abstract

Dirichlet regression, besides logratio regression, has been widely used in modelling compositional data. In this paper, several approaches to compute residuals for Dirichlet regression are considered. The graphical and quantitative properties of the proposed residuals are investigated. Finally, the techniques proposed are illustrated with two applications.

Key words: Aitchison distance, Compositional data, Dirichlet regression, pseudo residuals

1 Introduction

Compositional data arise commonly in many disciplines where scientists are interested in the dependence of non-negative proportions with unit-sum on certain relevant factors. For example, economists might be interested in how the composition of household income spent on food, clothes, entertainment and services is influenced by a set of family characteristics such as the total income, the size, and number of children. In biology, Campbell and Mosimann

(1987) were interested in modelling the shape composition of Mexican turtles and human body based on the total size and the body surface area respectively. Similarly, compositional data analysis arise in sedimentology (Coakley and Rust 1968), geochemistry (Aitchison 1984), archaeometry (Baxter and Freestone, 2006) and psychiatry (Gueorguieva et al. (2008)).

Due to the unit-sum constraint and its consequences, traditional regression models are not suitable for modelling such data. Aitchison (1986) suggested an analysis based on the logratios of the compositional data. Campbell and Mosimann (1987) developed an alternative approach by extending the Dirichlet distribution to a class of Dirichlet Covariate Models (Dirichlet Regression). Several approaches to model compositional data using generalized Liouville family have been introduced by Rayens and Srinivasan (1994), Smith and Rayens (2002), and Iyengar and Dey (2002).

Campbell and Mosimann (1987) and Hijazi (2003) developed maximum likelihood inference in Dirichlet regression models. Hijazi (2006) investigated the diagnostics checking in Dirichlet regression to assess the model validity and identify the outlying and influential compositions. Gueorguieva et al. (2008) have used Dirichlet regression in modelling psychiatric data and provided several definitions of residuals and overdispersion and influence diagnostics.

In this paper, we investigate the different forms of residuals in compositional data analysis and compare their use in diagnostics analysis in Dirichlet regression. Section 2 presents the Dirichlet regression model and four different residuals. Applications to illustrate the use of the proposed residuals are presented in Section 3. Finally, concluding remarks are given in Section 4.

2 Dirichlet Regression Residuals

Let y $=(y_1, ..., y_D)$ be a $1 \times D$ positive vector having Dirichlet distribution with positive parameters $(\lambda_1, ..., \lambda_D)$ with density function

$$f(\mathrm{y}) = \frac{\Gamma(\lambda)}{\prod\limits_{j=1}^{D} \Gamma(\lambda_j)} \prod_{j=1}^{D} y_j^{\lambda_j - 1} \tag{2.1}$$

where $\sum\limits_{j=1}^{D} y_j = 1$ and $\lambda = \sum\limits_{j=1}^{D} \lambda_j$.

In Dirichlet regression model, the parameters can be written as $\lambda_{ij} = h_j(x_i)$ for $j = 1, 2, ..., D$ where $h_j(x_i)$ is a positive-valued function of the covariate x. For example, in quadratic model $h_j(x_i)$ can be written as $h_j(x_i) = \beta_{j0} + \beta_{j1}x_i + \beta_{j2}x_i^2$. Under this parametrization, the mean and variance of the resulting distribution would be

$$\mu_{ij} = \frac{h_j(x_i)}{h(x_i)} \quad .$$

and

$$\sigma_{ij}^2 = \frac{\mu_{ij}(1 - \mu_{ij})}{1 + h(x_i)}.$$

where $h(x_i) = \sum\limits_{i=1}^{D} h_i(x)$. The Dirichlet regression coefficients $(\beta_{jk})$are then estimated using the maximum likelihood method (Hijazi 2003) and used to compute $\hat{\lambda}_{ij}$ and $\hat{y}_{ij}$.

Residuals analysis is useful for assessing the adequacy of the model assumptions and identifying atypical observations. Generally, the residuals represent the distances between the observed response and the fitted conditional mean. In compositional data, different definitions of the distance have been proposed yielding different forms of residuals.

Utilizing the probability integral transform and the relationship between Dirichlet and beta distributions, Hijazi (2006) proposed the following *pseudo residuals*

$$z_{ij} = \Phi^{-1}(F_j(y_{ij})) \tag{2.2}$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution and F_j is the cumulative distribution function of beta distribution with estimated parameters $\hat{\lambda}_{ij}$ and $\left(\sum\limits_{j=1}^{D} \hat{\lambda}_{ij}\right) - \hat{\lambda}_{ij}$. If the Dirichlet distribution is the correct model, the pseudo residuals, z_{ij}'s, follow the standard normal distribution and can be treated as standardized residuals in linear regression. The inspection of the pseudo residuals for individual components of compositions would be helpful in checking the model misspecification but not in identifying atypical compositions. Even though the multivariate residuals, $\mathbf{z} = (z_1, ..., z_D)$ do not follow a multivariate normal Mahalanobis distance might be used to identify the outlying and atypical observations and hence the corresponding outlying compositions.

In logratio regression, the residuals analysis is mainly carried out on the transformed compositions using the diagnostics in multivariate regression. Aitchison (2005) suggested the staying-in-the-simplex or *compositional residuals* defined as

$$\mathbf{r}_i = \mathbf{y}_i \ominus \hat{\mathbf{y}}_i = C\left(\frac{y_{i1}}{\hat{y}_{i1}}, ..., \frac{y_{iD}}{\hat{y}_{iD}}\right) \tag{2.3}$$

where C is the closure operation (Aitchison, 1986). If the model fits the compositional data well, these residuals should be spread around the center of a ternary diagram and then might be used to identify the potential outlying compositions (Aitchison 2005). The distribution of these residuals in Dirichlet regression is difficult to handle but our simulation studies yielded results consistent with logratio analysis. The simulation studies have also indicated that when the model is misspecified, the residuals are spread around a point far from the center of the ternary diagram or form trends rather than random scatter. A composition is considered atypical if the corresponding compositional residual falls far from the cluster in the center of the ternary diagram.

In the simplex geometry, Aitchison (1986) proposed *Aitchison distance* (Δ) as a measure of the distance between two compositions. This distance can be used as a measure of the distance between the observed composition and the corresponding fitted one as follows

$$\Delta(\mathbf{y}_i, \hat{\mathbf{y}}_i) = \left[\sum_{j=1}^{D}\left\{\log\frac{y_{ij}}{g(\mathbf{y}_i)} - \log\frac{\hat{y}_{ij}}{g(\hat{\mathbf{y}}_i)}\right\}^2\right]^{1/2} \tag{2.4}$$

where $g(\mathbf{w})$ is the geometric mean of the composition $\mathbf{w}$. The distribution of this distance under Dirichlet does not have a closed form which limits the use of these residuals in checking the model goodness of fit. However, the quantiles of such distribution can be easily estimated using bootstrap and used to single out atypical compositions.

Boyles (1997) introduced a modified chi-squared statistic as a measure of agreement between observed compositions and a target composition to use in multivariate control charts. Boyles showed that this statistic is asymptotically distributed as chi-square with $D-1$ degrees of freedom. The proposed statistic can work as a *Chi-square residual* in compositional data as follows

$$c_i = \left(\hat{\lambda}_i + 1\right) \sum_{j=1}^{D} \frac{\left(y_{ij} - \hat{y}_{ij}\right)^2}{\hat{y}_{ij}} \tag{2.5}$$

where $\hat{\lambda}_i = \sum_{j=1}^{D} \hat{\lambda}_{ij}$.

The resulting residuals can be tested to ensure the model specification and critical values can be computed to identify the outlying compositions with large chi-squared values.

It is noteworthy that residuals used in Beta regression (Ferrari, and Cribari-Neto(2004); Espinheira et.al. (2008)) can be extended to fit in the Dirichlet regression diagnostics tools. Gueorguieva et al. (2008) proposed several univariate and multivariate residuals to use in outlier detection. The univariate or marginal residuals can be used to identify the outliers at the component level while the multivariate residuals can be used to single out outlying compositions. The distributions of the proposed residuals have not been investigated making such residuals useless in assessing the goodness of fit of the model.

3 Applications

In this section, we present two applications to illustrate the ideas introduced in the paper. The first application employs simulated data while the second one is based on real data.

3.1 Simulated Data Example

The simulated data in Figure 1 consist of 100 compositions randomly generated from

$$\mathcal{D}(\beta_{10} + \beta_{11}x, \beta_{20} + \beta_{21}x, \beta_{30} + \beta_{31}x)$$

where the parameters used are $\beta_{10} = 5, \beta_{11} = 2, \beta_{20} = 10, \beta_{21} = -1, \beta_{30} = 2$ and $\beta_{31} = 0.5$ and the covariate values (x) were generated from the uniform distribution $U(0,5)$. To investigate the behavior of the proposed residuals in the existence of outlying compositions, we added the composition (0.25, 0.70, 0.05) to the simulated data where the covariate value is $x_{101} = 4$. The visual inspection of Figure 1, with this composition circled, does not indicate the atypicality of such observation.

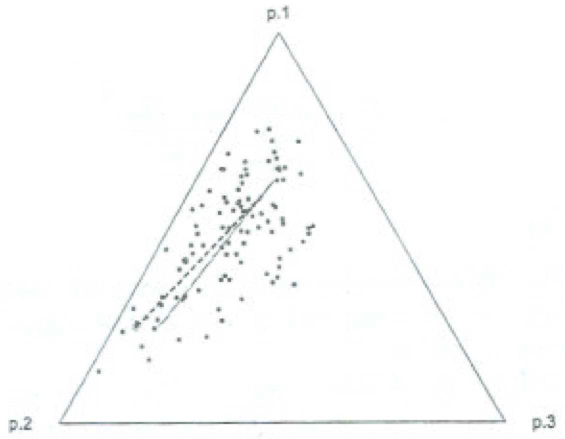

Figure 1: The simulated data: solid line represents the fitted model and dashed line represents the distance between the 101st compositions and its fitted value

The residuals from the fitted regression model with composition 101 are given in Figure 2. The Mahalanobis distance plot in Figure 2a singles out the 101st composition as outlier. Similarly, the plots of Aitchison distance and chi-square residuals, Figures 2c and 2d, single out the same composition as the only outlying one. The plot of the compositional residuals, with the 101st compositional residual circled, in Figure 2b does not clearly identify this composition as atypical. This is due to the large spread of the compositional residuals over the ternary diagram.

Table 1: Model fit: Simulated data

	Without obs. 101			With obs. 101		
Parameter	Estimate	Std. error	p-value	Estimate	Std. error	p-value
β_{10}	3.68	0.9929	0.0002	3.87	0.9912	0.0001
β_{11}	3.48	0.5449	0.0000	2.80	0.6037	0.0000
β_{20}	10.22	1.4143	0.0000	9.68	1.4729	0.0000
β_{21}	−0.66	0.3735	0.0908	−0.73	0.3878	0.0499
β_{30}	1.51	0.3717	0.0001	1.56	0.3758	0.0000
β_{31}	0.90	0.1759	0.0000	0.71	0.1929	0.0001

Table 1 gives the parameter estimates, the standard errors and the p-values for the corresponding significance tests for the model with and without

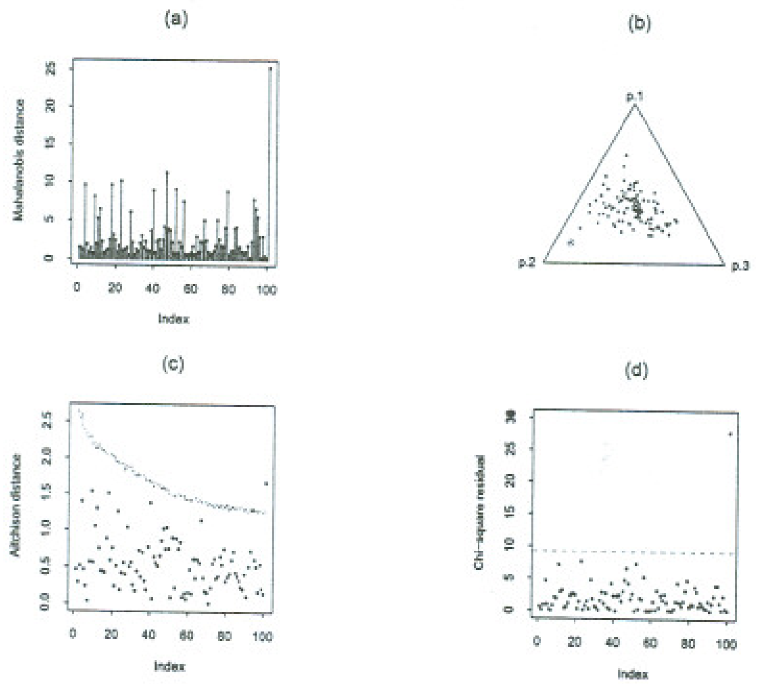

Figure 2: The residual plots for the simulated data: (a) Mahalanobis distance of pseudo residuals (b) Compositional residuals (c) Aitchison distance (d) Chi-square residuals

composition 101. The removal of this composition has clear impact on the estimates of the slopes β_{11}, β_{21} and β_{31} but not on the overall estimated model. Figure 1 gives the fitted model with and without composition 101. The two fitted lines almost coincide indicating no significant effect of this composition on the model fit. The dashed line in the ternary diagram connects composition 101 with its fitted value showing the large distance between the two compositions. This is an indication of the atypicality of this composition.

3.2 Real Data Example

The second application concerns data on human body surface area of 34 individuals aged six hours to sixty years (Campbell and Mosimann, 1987). The knowledge of the proportional body surface area is a useful measure in

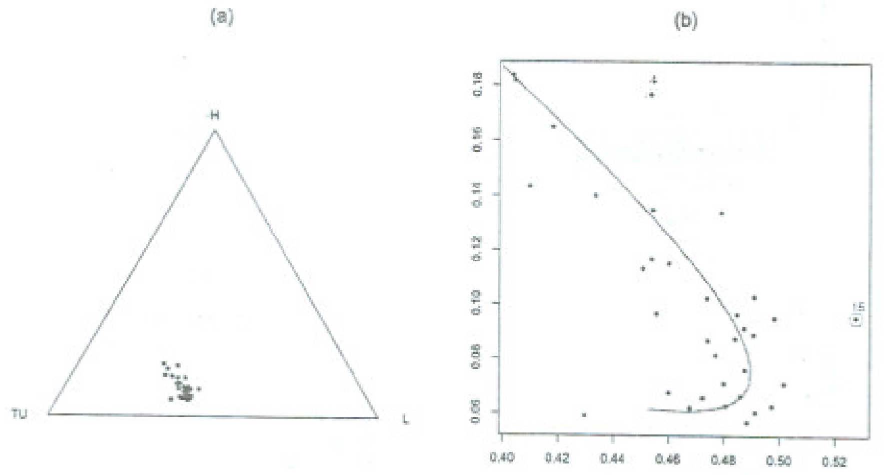

Figure 3: Distribution of the Human Body Surface Area: (a) Data in the simplex (b) A zoom out of the data in the simplex

burn therapy. The total body surface area varied from 0.206 to 2.036 square meters. The data consist of the proportions of the head (H), trunk and upper extremities (T+U) and lower extremities (L). It is believed that the proportional body area depends on the total body surface area (x). The ternary diagram in Figure 3 shows the distribution of the proportional data. The trend of the proportional change shows a J-shaped behavior of the proportions in the simplex suggesting a quadratic model of the following form

$$\mathcal{D}(\beta_{10}+\beta_{11}x+\beta_{12}x^2, \beta_{20}+\beta_{21}x+\beta_{22}x^2, \beta_{30}+\beta_{31}x+\beta_{32}x^2)$$

The residuals plots of the fitted model are given in Figure 4. The Mahalanobis distance plot of the pseudo residuals (Figure 4a) singles out composition 4 as outlier. However, the Mahalanobis distance plot (Figure 4d) identifies composition 15 as potential outlier which falls above the 99% critical limit in Figure 4c however composition 4 will be classified as outlier at 95% level. The plot of Aitchison distance (Figure 4c) does not detect any

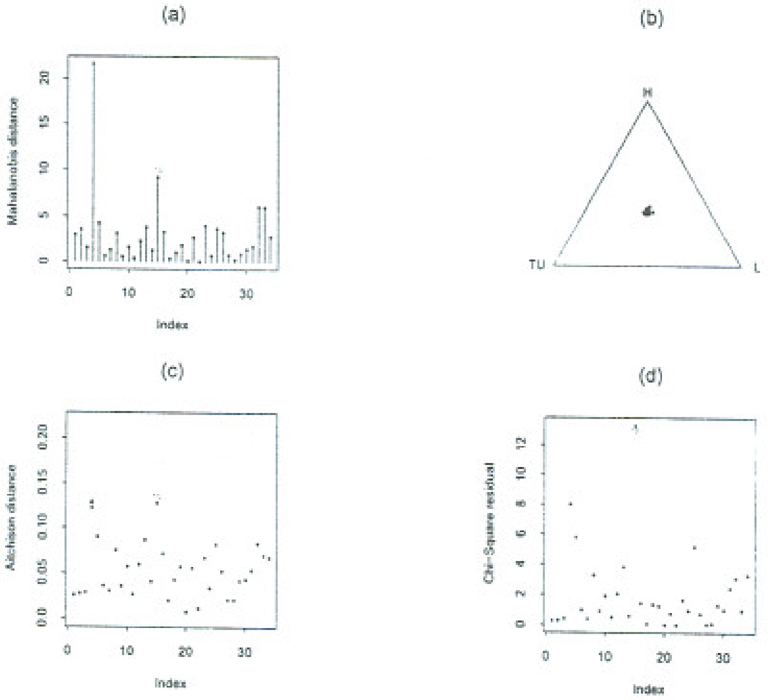

Figure 4: The residual plots for the surface area data: (a) Mahalanobis distance of pseudo residuals (b) Compositional residuals (c) Aitchison distance (d) Chi-square residuals

outlier at 99% level. The compositional residuals plot (Figure 4b) shows that the residuals are tightly spread around the center of the ternary diagram. The plot does not single any atypical observations. The model fit in Table 2 indicated that β_{12} is the only significant quadratic parameter. The model shows a good fit for the surface proportions in the simplex. The fitted model follows the J-shaped curvature of the compositions. The estimated parameters after removing compositions 4 and 15 are given in Table 2. The model fit did not change substantially where the two models almost coincide in the ternary diagram. The quadratic term in the lower extremities component, β_{32}, became significant besides β_{12}. The residuals plots (not shown here) did not show any signs of lack of fit nor identified atypical compositions. It is worth mentioning that the extreme point in Figure 3 at the lower end of the J-curve corresponds to an obese adult with the largest body surface

Table 2: Model fit: Body Surface Area Data

	Full data			Without obs. 4 and 15		
Parameter	Estimate	Std. error	p-value	Estimate	Std. error	p-value
β_{10}	391.65	85.11	0.0000	628.86	129.18	0.0000
β_{11}	−292.67	84.25	0.0005	−454.02	114.75	0.0001
β_{12}	76.72	28.60	0.0077	98.37	34.95	0.0049
β_{20}	1021.45	231.61	0.0000	1638.77	349.50	0.0000
β_{21}	−89.71	306.95	0.7701	88.94	381.56	0.8157
β_{22}	51.71	196.49	0.7924	−183.83	234.29	0.4327
β_{30}	359.06	97.38	0.0002	569.50	139.19	0.0000
β_{31}	490.12	218.11	0.0246	910.19	294.10	0.0020
β_{32}	−168.36	138.68	0.2247	−428.24	174.93	0.0144

area (over 2 square meters). The deletion of this point dose not have any significant effect on the model fit.

4 Concluding Remarks

In this paper, we have investigated the residual analysis in Dirichlet regression. An overview of the proposed residuals in compositional data analysis is provided. The individual residuals plots of the pseudo residuals can be used to identify model misspecification while the Mahalanobis distance is useful identifying the atypical compositions. The chi-square residuals plot is useful in assessing the goodness of fit of the model and identifying the atypical compositions. The inspection of Aitchison distance, with the simulated quantiles, helps only in identifying the outlying compositions. The ternary plot of the compositional residuals will be helpful in identifying model misspecification but closer look at the distribution of such residuals might be helpful in identifying atypical compositions. The applications presented illustrated the uses of the proposed residuals in the assessment of Dirichlet models.

References

[1] Aitchison, J. (1984). The statistical analysis of geochemical compositions. *Mathematical Geology*, 16(6), 531564.

[2] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, New York.

[3] Aitchison J., (2005). *A Concise Guide to Compositional Data Analysis.* CDA Workshop, Girona.

[4] Baxter, M J and Freestone, I C (2006). Log-ratio compositional data analysis in archaeometry. *Archaeometry*, 48, 511-531.

[5] Boyles, R. (1997). Using chi-square statistic to monitor compositional process data. *Journal of Applied Statistics*, 24(5), 589-602.

[6] Campbell, G., and Mosimann, J. E. (1987). Multivariate methods for proportional shape. *ASA Proceedings of the Section on Statistical Graphics*, 10-17.

[7] Coakley, J.P. and Rust, B.R. (1968). Sedimentation in an Arctic lake. *Journal of Sedimentary Petrology*, 38, 1290-1300.

[8] Espinheira, P. L., Ferrari, S. L. P., Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, 35(4), 407419.

[9] Ferrari, S. L. P., Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, 31(7), 799815.

[10] Gueorguieva, R., Rosenheck, R., Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data . *Computational Statistics and Data Analysis*, 52(12), 5344-5355.

[11] Hijazi, R. (2003). Analysis of Compositional data using Dirichlet Covariate Models. *PhD Dissertation*, The American University, Washington, DC.

[12] Hijazi, R. (2006). Residuals and Diagnostics in Dirichlet Regression. *ASA Proceedings of the Joint Statistical Meetings 2006, American Statistical Association*, 1190-1196.

[13] Iyengar, M., and Dey, D. (2002). A semiparametric model for compositional data analysis on the simplex. *Test*, 11, 303-315.

[14] Rayens, W. and Srinivasan, C. (1994). Dependence properties of generalized Liouville distributions on the simplex. *Journal of the American Statistical Association*, 89, 14651470.

[15] Ronning, G. (1989). Maximum likelihood estimation of Dirichlet distributions. *Journal of Statistical Computation and Simulation*, 32, 215-221.

[16] Smith, B. and Rayens, W. (2002). Conditional generalized Liouville distributions on the simplex Statistics. *Statistics*, 36 (2), 185-194.