# Chapter 3. Model validation for MLE logistic regression on cyclic data using bootstrapping.

# Abstract

Motivated by common chi-squared assumptions made for the modelling of logistic regressions, we investigate the distribution of logistic regression parameters and model goodness-of-fit for binary data with cyclic sampling properties using the bootstrap. Two re-sampling methods for the bootstrap are demonstrated, illustrating and compensating for the effects which sparse sampling can have on the distributions of the model fit and logistic parameters. For non-sparse data, we demonstrate the inadequacy of the chi-squared assumption for the logistic model fit, whilst providing more realistic and robust bootstrapped confidence intervals for the logistic model parameters.


# 1 Introduction

In this paper, methods for estimating binomial variables with cyclic properties are investigated using the parametric bootstrap. The motivation for developing this method was provided by Upton et al. (2003) who developed a method of cyclic logistic regression using maximum likelihood estimation (MLE) for modelling this type of data. They based confidence intervals for the model parameters on the model deviance using a chi-squared distribution with one degree of freedom, and based a goodness-of-fit test on a chi-squared distribution with degrees of freedom equal to the number of sampling orientations minus the number of model parameters. However, a $\chi^2$ distribution for a model is only exact for normal (Gaussian) models and, in general, does not provide a very good approximation for the sampling distribution (Dobson 1990). To better approximate the distribution of the model fit, we utilise the parametric bootstrap.

The bootstrap approach can have several complications arising from sampling inadequacy (Efron & Tibshirani 1993); however, once this issue is overcome, more robust and realistic estimates can be derived for model parameters and overall model fit. Here we illustrate a method of employing bootstrapping for the MLE of the cyclic logistic regression to obtain robust estimates of the variability of the model parameters and the goodness-of-fit of the model. Examples from geological settings are used.

# 2 Geological Background

The geological problem of measuring the trend of a linear fabric in rocks, the foliation intersection or inflection axis (*FIA*) preserved within porphyroblasts, was examined by Upton et al. (2003) and is revisited here. FIA are the axes of curvature of surfaces (*inclusion trails*) overgrown by and preserved inside of mineral grains (*porphyroblasts*) that are large compared to the smaller sizes of grains in the surrounding matrix (Fig. 1). It has been proposed that FIA record the orientation of fold axes that existed when the porphyroblasts grew (Hayward 1990).

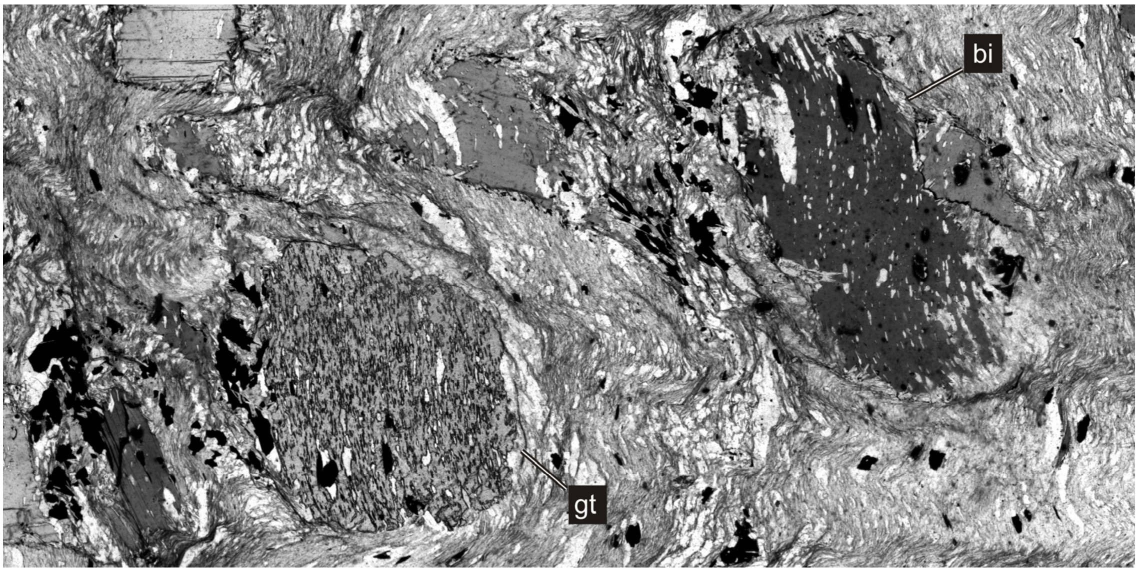


**Figure 1.** This is a photomicrograph of a vertical thin section of schistose rock containing garnet (gt) and biotite (bi) porphyroblasts. Both the garnet and biotite contain inclusion trails, made up of elongate quartz grains, which have an anticlockwise or "Z" shaped asymmetry. The foliation in the matrix is defined by the alignment of quartz and mica and is approximately vertical. The bottom edge of the image is approximately 8mm long.

Furthermore, the orientation of these axes may reflect the relative motion of the tectonic plates that form the Earth's crust (Bell et al. 1995). These axes are of considerable interest to geologists and methods for their statistical analysis are important.

The FIA orientation in a single porphyroblast reflects that of the foliations (sub-planar alignment of matrix grains) in the rock that the inclusion trails preserve. Their orientations vary between porphyroblasts within a rock sample. FIA are three-dimensional data with a spherical distribution; however, geologists are most interested in their trend in a horizontal plane reducing the problem to two dimensions. It is generally not possible to measure the FIA in individual porphyroblasts, so the method used provides an estimate of the mean trend of the FIA in a given rock sample. The asymmetry method relies on the fact that a simple asymmetrically folded surface with a sub-horizontal axis will appear to have opposite asymmetries when cut by two vertical planes that strike either side of the fold axis (Bell et al. 1998). The fold asymmetry appears anticlockwise in one ("Z", left side of Fig. 2a) and clockwise in the other ("S", right side of Fig. 2a). Note that both these planes are being viewed in the same direction – clockwise about a vertical axis. Curved inclusion trails preserved in porphyroblasts are analogous to such a fold. Figure 2b-h shows how this concept is applied to measuring FIAs. Geologists prepare thin slices of rock (*thin sections*) that can be examined with an optical microscope. The asymmetries of inclusion trails in porphyroblasts intersected by the thin section are recorded. The FIA trend is recorded as being the midway point between the two vertical thin-sections with a close

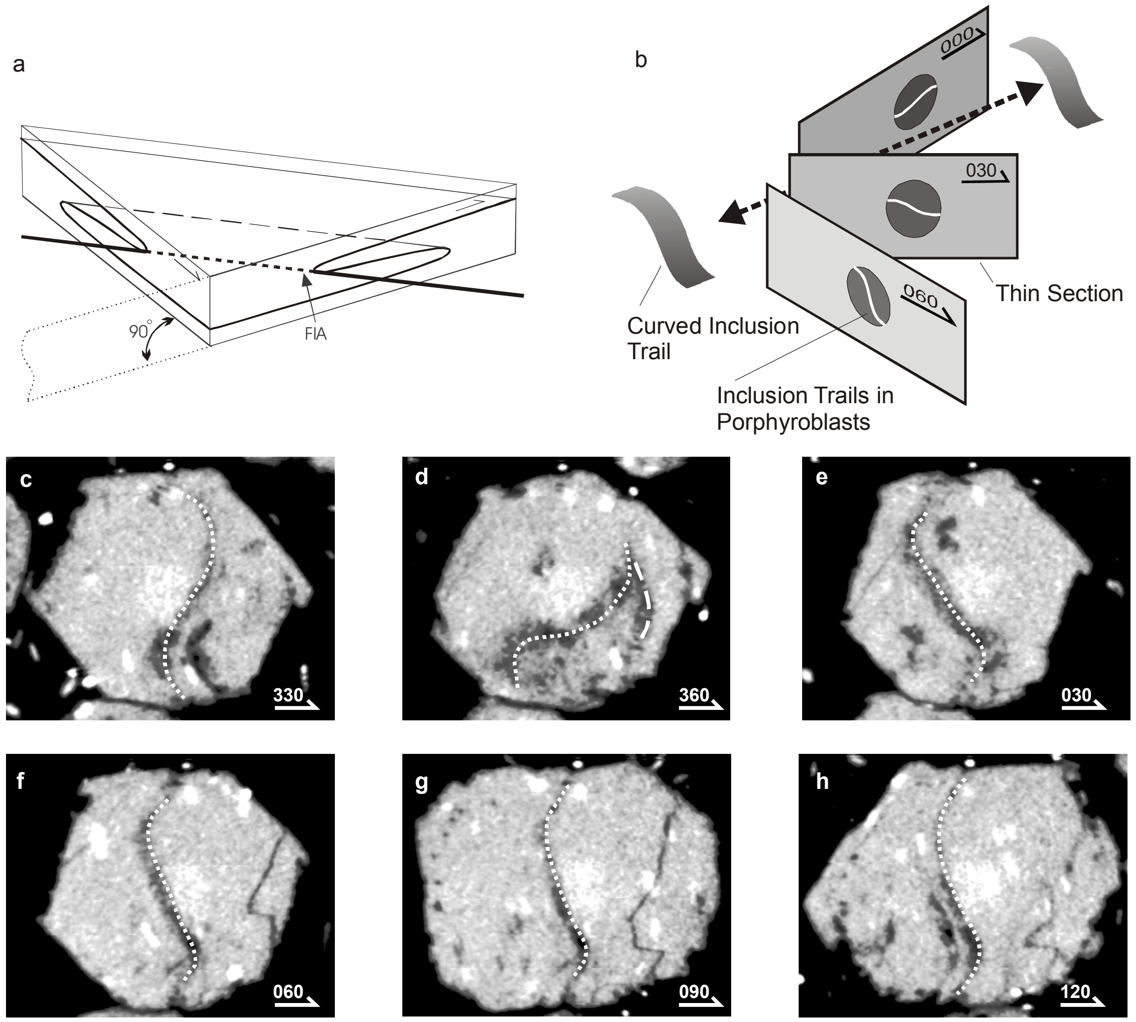


**Figure 2.** An illustration of what FIA are and the asymmetry method used to determine them. A slab of "rock" with an asymmetrically folded surface is shown in (a), cut by two vertical sections 90º apart. The fold asymmetry appears anticlockwise in the section one on the left ("Z" shaped) and clockwise in the section on the right hand side ("S" shaped). The FIA must lie between these sections (modified from Bell et al. 1998). (b) shows how the asymmetry method is applied in thin sections. Three sections are shown spaced at 30º, with the FIA axis between the 0º and 30º sections. The dark grey circle represents a porphyroblast with the inclusion trail asymmetry shown in white, flipping from clockwise in the 0º to anticlockwise in the 30º and 60º sections. An example from the V209 sample is shown in (c) to (h). These are sections at 30º radial increments through X-ray CT data for a single porphyroblast (light grey). The dotted white lines highlight the inclusion trails; they switch asymmetry from anticlockwise to clockwise between the 360º and 30º sections.

angular spacing where the observed dominant asymmetry switches. Sections are typically cut at a 10˚ angular spacing in the vicinity of the FIA.

High Resolution X-ray Computed Tomography (HRXCT) has been used to measure the FIA in 58 porphyroblasts from a single sample in chapter 1. A set of asymmetry observations was derived from this dataset in order to test the MLE cyclic logistic regression. The asymmetry was determined in each individual porphyroblast sectioned at 10 ° increments from 0° to 170°

| ***Orientation*** | ***Clockwise*** | ***Anticlockwise*** | ***Total*** |
|---|---|---|---|
| 0 | 2 | 56 | 58 |
| 10 | 11 | 47 | 58 |
| 20 | 24 | 34 | 58 |
| 30 | 50 | 8 | 58 |
| 40 | 56 | 2 | 58 |
| 50 | 57 | 1 | 58 |
| 60 | 58 | 0 | 58 |
| 70 | 58 | 0 | 58 |
| 80 | 58 | 0 | 58 |
| 90 | 58 | 0 | 58 |
| 100 | 58 | 0 | 58 |
| 110 | 58 | 0 | 58 |
| 120 | 58 | 0 | 58 |
| 130 | 58 | 0 | 58 |
| 140 | 58 | 0 | 58 |
| 150 | 58 | 0 | 58 |
| 160 | 58 | 0 | 58 |
| 170 | 57 | 1 | 58 |

**Table 1** Asymmetry data based on 58 FIA measurements from HRXCT data from sample V209.

and the results are shown in table 1. Deriving the asymmetry data in this manner simulates the way that it is normally collected using the asymmetry method. In addition, the data used by Upton et al. (2003) is also re-examined.

# 3 Cyclic Logistic Regression

Measurements of FIA orientations are characterised by asymmetry counts on an angular region of periodicy of π. Further symmetry of the counts occurs through a reflection about the axis of the FIA. If we define μ as the direction of the FIA and $\theta_i$ as the $i^{th}$ orientation from north, then the following properties apply regarding the probability, $P(\theta_i) = \varpi_i$, of observing a clockwise inclusion trail geommetry as opposed to an anti-clockwise one:

$$P(\theta_i - \mu) = P(\theta_i - \mu + \pi) = 1 - P(\mu - \theta_i) \tag{1}$$

with $0 \le \mu < \pi, 0 \le \theta_i < \pi$ and $P(\theta_i - \mu)$ monotonically decreases (or increases) from $P(\pm\pi/2) = 1$ to $P(\mp\pi/2) = 0$ with $P(0) = 1/2$. Thus, the probability of observing a positive asymmetry (clockwise) at an angle $\theta_i$ is symmetric and cyclic about $\mu$. Following Upton et al. (2003), the logistic regression with a sinusoidal transform to model $P(\theta_i - \mu)$ is used to model π cyclic effects.

$$x_i = \sin(\theta_i - \mu) \tag{2}$$

The logistic function is then used to model the probability of observing a particular asymmetry.

$$P(x_i) = \frac{1}{1+e^{\beta x_i}} \tag{3}$$

# 4 Maximum Likelihood Estimation

Given a specified model, $P(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ with parameters $\boldsymbol{\phi}$, observed data, $\boldsymbol{\theta}$ and joint probability density function $L(\boldsymbol{\phi} \mid \boldsymbol{\theta})$, the maximum likelihood estimator, $\hat{\boldsymbol{\phi}}$, is defined as

$$L(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\theta}) \geq L(\boldsymbol{\phi} \mid \boldsymbol{\theta}); \forall \boldsymbol{\phi} \in \Theta, \tag{4}$$

where $\Theta$ represents all possible values of $\boldsymbol{\phi}$. Given $L(\boldsymbol{\phi} \mid \boldsymbol{\theta})$ the data is treated as fixed and $\boldsymbol{\phi}$ as a variable parameter(s). The value for $\hat{\boldsymbol{\phi}}$ that maximises the joint probability of the data is sought.

For count type data the model can generally be defined as $P(\boldsymbol{\theta}_\iota \mid \boldsymbol{\phi}) = \varpi_i$, where $\varpi_i$ is the probability of a success (e.g. clockwise observation) on the $i^{\text{th}}$ orientation. If there are $r_i$ successes out of a total number observations at the $i^{th}$ position, $n_i$, then the joint probability function is (Dobson 1990)

$$L(\boldsymbol{\phi} \mid \boldsymbol{\theta}) = \prod_{i=1}^{m} \varpi_i^{r_i} (1-\varpi_i)^{n_i - r_i}; \quad i = 1 \ldots m \tag{5}$$

Equally, the value $\hat{\boldsymbol{\phi}}$ that maximises Eqn. 5 is also the value that maximises the log-likelihood function so that $\ell(\boldsymbol{\phi} \mid \boldsymbol{\theta}) = \log(L(\boldsymbol{\phi} \mid \boldsymbol{\theta}))$ since the logarithmic function is monotonic. Thus

$$\ell(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\theta}) \geq \ell(\boldsymbol{\phi} \mid \boldsymbol{\theta}); \forall \boldsymbol{\phi} \in \Theta \tag{6}$$

and the equivalent expression for Eqn (5) is

$$\ell(\boldsymbol{\phi} \mid \boldsymbol{\theta}) = \sum_{i=1}^{m} \{ r_i \log(\varpi_i) + (n_i - r_i) \log(1-\varpi_i) \} \qquad i = 1 \ldots m \tag{7}$$

Commonly, it is easier to work (optimise) the log-likelihood function than the likelihood function itself. For the above cyclic logistic regression model in Eqn. (3), the log likelihood function is

$$\ell(\beta, \mu \mid \theta) = \sum_{i=0}^{m} r_i \sin(\theta_i - \mu)\beta - n_i \left( \sin(\theta_i - \mu)\beta + \ln\left(1 + e^{\beta \sin(\theta_i - \mu)}\right) \right) \tag{8}$$

where $r_i$ is the number of clockwise orientations out of a total number of observations, $n_i$, at orientation $\theta_i$. The maximum of Eqn. (4) corresponds to the most likely values $\left(\hat{\beta}, \hat{\mu}\right)$ given the data $\boldsymbol{\theta}$.

Using MLE, an estimate of both the distribution of the model parameters and the goodness-of-fit can be determined based on a measure called deviance. Deviance is calculated as the relative change in the log-likelihood value with reference to a null hypothesis, which is usually the maximised model for parameter distributions or the saturated model for goodness-of-fit estimation. Typically, the deviance, D, is *assumed* to asymptotically follow a $\chi_p^2$ distribution with *p* degrees of freedom. Here *p* is the difference between the number of orientations and the number of model parameters.

For the goodness-of-fit,

$$^{G}D = 2\left(\ell^*\left(\boldsymbol{\theta}\right) - \ell\left(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\theta}\right)\right) \tag{9}$$

where $\ell^*\left(\boldsymbol{\theta}\right)$ is the saturated model. In the case of the cyclic logistic regression the model is given by

$$\ell^*\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{m} \left\{ r_i \log\left(r_i\right) + \left(n_i - r_i\right) \log\left(n_i - r_i\right) - n_i \log\left(n_i\right) \right\}. \tag{10}$$

$^{G}D$ is assumed to follow a $\chi_p^2$ distribution with $p = m - 2$ degrees of freedom. This assumes that under the null hypothesis the *m* orientations are independent with a standard deviation equal to 1 (Dobson, 1990).

The distribution of the model parameters, $^{M}D$, is given by

$$^{M}D = 2\left(\ell\left(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\theta}\right) - \ell\left(\boldsymbol{\phi} \mid \boldsymbol{\theta}\right)\right). \tag{11}$$

$^{M}D$ is assumed to follow a $\chi_p^2$ distribution with $p = 1\, or\, 2$ degrees of freedom, depending on whether either $\beta\, or\, \mu$ are fixed $\left(p = 1\right)$ or allowed to vary simultaneously $\left(p = 2\right)$.

The maximization of $\ell\left(\boldsymbol{\phi} \mid \boldsymbol{\theta}\right)$ is the equilivant of the minimization of $-\ell\left(\boldsymbol{\phi} \mid \boldsymbol{\theta}\right)$. We employed the quasi-Newton optimizer in Matlab to perform this operation.

# 5 Bootstrap

The fundamental principle of the bootstrap is that a resample *with replacement* of a random sample from a certain population is considered to be a random sample of that population (Efron & Tibshirani 1993). Thus, the collected data can be re-sampled multiple times and empirical distributions of population parameters (such as the mean) can be obtained, rather than relying on asymptotic assumptions about the data.

In the classical non-parametric bootstrap by Efron and Tibshirani (1993), re-sampling of the original sample, $\boldsymbol{\theta}^0$, is done with replacement such that the bootstrapped sample, $\boldsymbol{\theta}^j$, is the same size as the original data. That is, if there are *k* observations in $\boldsymbol{\theta}^0$, then $\boldsymbol{\theta}^0$ is sampled *k* times with replacement to form $\boldsymbol{\theta}^j$. In total, *B* bootstrapped samples are formed from which empirical distributions of the population parameters can be derived. In the case of MLE cyclic logistic regression, the distributions $\hat{F}_\beta(\theta), \hat{F}_\mu(\theta)\, and\, \hat{F}_{^G D}$ for the cyclic logistic regression parameters $\beta, \mu$ and $^G D$ respectively are important.

For each of the *B* bootstrapped samples $\boldsymbol{\theta}^j, j = 1 \ldots B$, MLE is used to calculate $\beta^j, \mu^j$ and $^G D^j$ using Eqns. (8-10), where $r_i$ and $n_i$ are from $\boldsymbol{\theta}^j$. Using these *B* estimates, the empirical distributions, $\hat{F}_\beta(\theta_i), \hat{F}_\mu(\theta_i)\, and\, \hat{F}_{^G D}$, are formed and from them the respective confidence intervals for $\beta, \mu$ can be found.

There are three primary methods that can be used to calculate confidence intervals (CI) using bootstrap methods: (1) CI based on *t*-tables; (2) the percentile method; and (3) the percentile method adjusted for bias and acceleration ($BC_a$). In this paper we provide CI based on the percentile and $BC_a$ methods as they tend to be more robust CI estimates (Efron & Tibshirani 1993).

To summarize, the percentile method sorts the bootstrapped values in ascending order and selects the $100 \cdot \alpha^{th}, \alpha \in (0,1)$, interval, $\left[\phi_L^{\alpha/2}, \phi_H^{\alpha/2}\right]$, such that $1-\alpha/2$ percent of the values are above and below the CI. Let $\hat{C}$ be the cumulative distribution function of $\hat{\phi}^j$. Thus

$$\hat{C}(\phi) = \int_{-\infty}^{\phi} \hat{F}(t)\, dt \tag{12}$$

and the $1-\alpha/2$ percentile interval is

$$\left[\phi_L^{\alpha/2}, \phi_H^{\alpha/2}\right] = \left[\hat{C}^{-1}(\alpha/2), \hat{C}^{-1}(1-\alpha/2)\right] \tag{13}$$

The BCa method is similar to the percentile method; however, a correction to the upper and lower $\alpha$ values, $\alpha^1, \alpha^2$, is made to allow for bias and acceleration in $\hat{F}_i(t_i)$ from a normal distribution. The values for $\alpha^1, \alpha^2$ are given by

$$\alpha^1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1-\hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right)$$
$$\alpha^2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)}\right) \tag{14}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and $z^{(\alpha)}$ is the $100\cdot\alpha^{th}$ percentile point of a standard normal distribution. The value $\hat{z}_0$ is a bias correction term approximated by the bootstraped samples:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\hat{\phi}^j < \hat{\phi}^0\right\}}{B}\right) \tag{15}$$

where $\hat{\phi}^0$ is the MLE from the original data. In addition, in Eqn (14), the term $\hat{a}$ corresponds to the acceleration, which for one parameter estimate, can be approximated by $\hat{z}_0$ to the second order.

Both of the above CI methods, and bootstrapping in general, are the result of the central limit theorem where $\hat{F}_\beta(\theta_i)$ and $\hat{F}_\mu(\theta_i)$ are approximately normal as *n,* the total sample size of the original data, gets large. This also raises the question of adequate sampling since if *n* is too small, the bootstrap method will produce biased estimates for $\hat{F}_\beta(\theta_i)$ and $\hat{F}_\mu(\theta_i)$ (Efron & Tibshirani 1993).

In this investigation, two methods for re-sampling are considered. The first resampling method (method A) consists of resampling on all observations irrespective of the orientation, while the second method (method B) resamples within each orientation. For example, if our data consisted of two orientations with 10 and 20 observations in each orientation $n_1$ and $n_2$ respectively, then for method A, each bootstrap sample will have 30 observations, however $n_1$ may or may not equal 10. For method B, $n_1$ will equal 10 since for the bootstrap sample the resampling is performed on each orientation individually.

The two bootstrap methods were investigated to determine the effect that sparse data would have. Method (A) treats all observations equally and can result in bootstrap samples that have no data in some orientations. In some cases, as in those provided by Upton et al. (2003) where for most orientations there is only one observation per orientation, the data is sparse. In these situations, the variability in the bootstrap samples using method A is largely due to the variation in the probability that the orientation(s) are sampled and not the variation of binomial probability at each orientation. Thus the distributions of $\beta$ and $\mu$ are more reflective of how the model parameters vary as the sampling distributions of the orientations vary near the FIA. Method (B) eliminates this effect by ensuring each orientation is represented in the bootstrap samples. However, for very sparse data, it may not be possible to employ method B since redrawing one observation from a sample size of one (per orientation) will result in the exact same sample and thus no variation in the cyclic logistic regression parameters.

For data rich samples, where the binomial variability can be estimated through resampling, the variation in the cyclic logistic regression parameters resulting from method A or B will be largely influenced by the binomial variation, rather than the distribution of sampling orientations.

# 6 Results and Discussion

Table 2 shows the results of the analysis of the data presented in Upton et al. (2003) using the chi-squared method and the two bootstrap methods. The confidence intervals for all methods exhibit broad similarities for $\beta$ and $\mu$, with the exception of the CA1 Rim and CA1 Core data. In the CA1 Rim, the CI from the method A bootstrap is substantially wider than that from the other methods. Due to the lack of sampling within orientations for the CA1 Core data, method B only generates one unique bootstrap sample, and hence results in a nil variation in $\beta$ and $\mu$.

| | | | CA1 | Rim | Middle | Core | Rim* |
|---|---|---|---|---|---|---|---|
| | | | *n* | 16 | 27 | 21 | 16 |
| | | | *m* | 10 | 14 | 12 | 10 |
| | | | $\mu$ | 49 | 19 | 74 | 50 |
| | | | $\beta$ | 92.13 | 3.48 | 3.54 | ~168 |
| | | | $^{G}D$** | 0 | 8.5 | 17.72 | 0 |
| *99% Confidence intervals* | | *Upton* | $\mu$ | (47,51) | (351,57) | (37,112) | (28,67) |
| | | | $\beta$ | - | (1.13,8.41) | (0.67,9.24) | (2.53,>800) |
| | *Bootstrap Method A* | *BCa* | $\mu$ | - | (5.4,84.0) | (33.6,89.6) | (34.8,84.4) |
| | | | $\beta$ | - | (0.92,111) | (-0.41,451) | (115,781) |
| | | *Percentiles* | $\mu$ | - | (355,63.5) | (40.2,97.7) | (26.1,70.8) |
| | | | $\beta$ | - | (1.51,760) | (1.55,685) | (88.2,634) |
| | *Bootstrap Method B* | *BCa* | $\mu$ | - | (10.5,52.2) | NA | (45.1,49.7) |
| | | | $\beta$ | - | (1.72,118) | NA | (133,132) |
| | | *Percentiles* | $\mu$ | - | (0,44.8) | NA | (45.1,56.2) |
| | | | $\beta$ | - | (1.77,776) | NA | (132,577) |

* Results for MLE and CI for Rim data reprocessed using more accurate maximization techniques than those used in Upton et al. (2003).
** $^{G}D$ is equivalent to Upton's "C".

**Table 2.** Results for the CA1 Core, Middle and Rim data for the data presented in Upton et al. (2003). Confidence intervals for model parameters could not be calculated for the core data using bootstrap method B as only one permutation of the data is possible. In all cases 5000 bootstrap resamples were made.

The Core, Mid and Rim data are all sparsely sampled. This has produced three unique results with bootstrapping and the Chi-Squared approximation. To assess the model validity, we examine the empirical distributions of $\beta$ and $\mu$ and the model deviance in Figures 3 and 4, which, theoretically, should be approximately normal for $\beta$ and $\mu$ and $\chi^2$ for the model deviance. Excessive deviation from the normal distribution is evident and is a result of inadequate sampling. For method A, the bootstrap on this sparse data is not adequately modelling the variability in the binomial probability at each orientation; it is modeling the probability that the orientation(s) are sampled instead. Thus the distributions of $\beta$ and $\mu$ are more reflective of how the model parameters vary as the sampling distributions of the orientations vary. Method B will not be affected in this way, as the probability of an orientation being sampled is invariant. However, it is affected by the sparse nature of the data in that a small number of unique data permutations are generated from the 5000 bootstrap samples (4, 60 and 1 for the CA1 Rim, Mid and Core data respectively).

In the case of the CA1 Mid data, the relative frequencies of the 60 permutations tend towards a normal distribution for $\mu$; the distribution for $\beta$ is also approaching a more central distribution (see inset in Fig. 4b) although it has outliers that distort it. The deviance shows a limiting $\chi^2$ distribution. The CA1 Core data shows the opposite extreme with the paucity of the data resulting in only one possible permutation, therefore acting as a point estimator for $\beta$ and $\mu$. Only the method B bootstrap of the CA1 Mid data demonstrates an adequate model fit.

Table 3 shows the results for the combined data (Upton et al. 2003) and the v209 data. There is good agreement between all three methods and the model parameters clearly exhibit asymptotic normality (Figure 5); this demonstrates both adequate data sampling and model fit (see also figure 6a-b). However, there is one main discrepancy in assessing the model fits between the bootstrapping procedures and the $\chi^2$ approach; this is the determination of the appropriate degrees of freedom of the model deviance. The degrees of freedom of the model deviance for the v209 and combined data are estimated to be approximately 5 and 12 respectively (Fig. 5). According to the classical $\chi^2$ approach, the degrees of freedom would be 16 and 15 ($m$-2) respectively.

The discrepancy in the degrees of freedom for the model fit can be attributed to dependence of the log-likelihood deviance on variation at each orientation. In Upton (2003), the assumption of unit independence of the sampling orientations in the logistic model gives rise to the approximation of a $\chi^2$ fit with m-2 degrees of freedom. This assumption of independence is not the case here, as the correlations of deviance between orientations are substantially different from zero (Table 4). If there were independence between orientations,

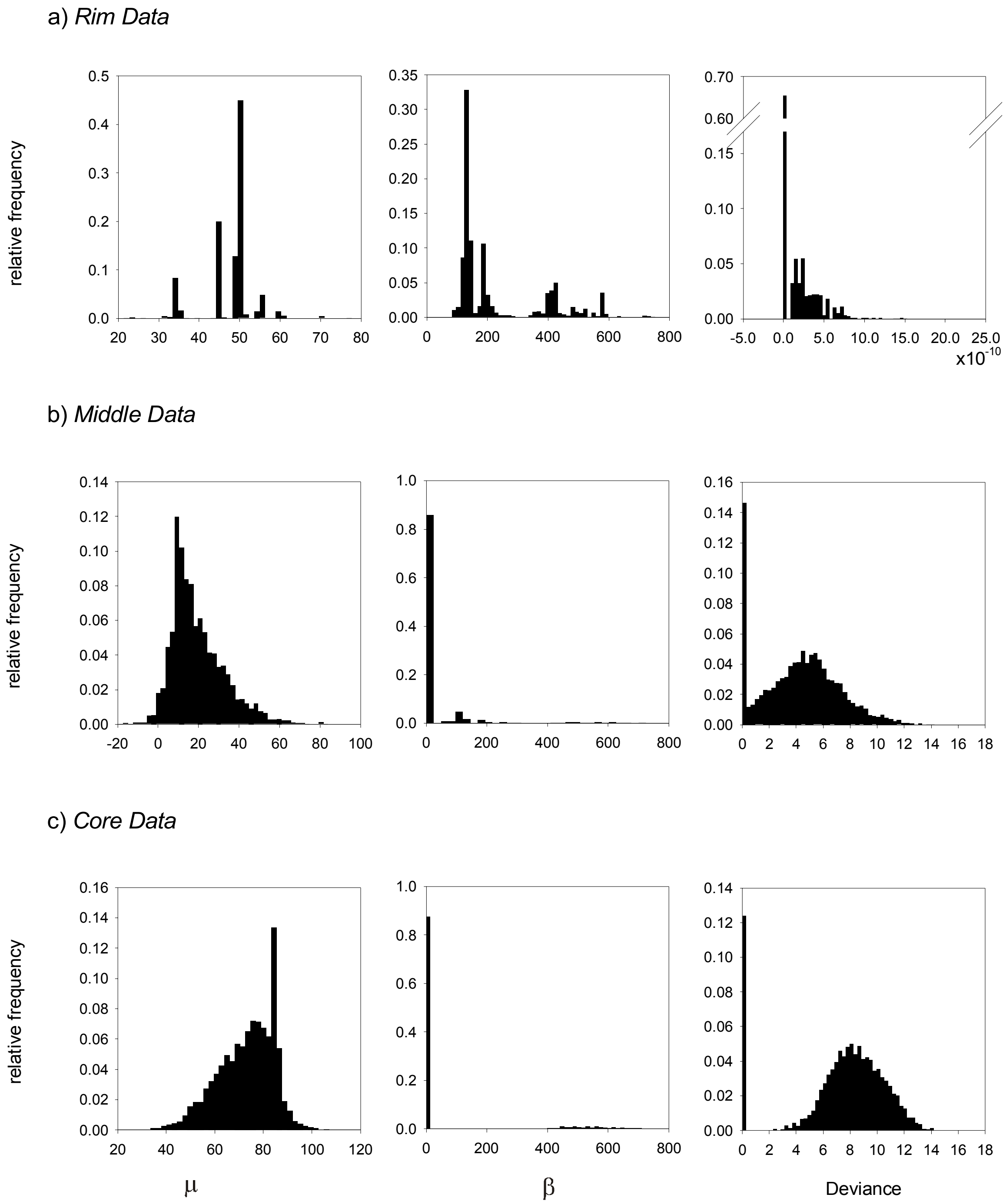


**Figure 3.** Bootstrapped empirical distributions of $\beta$, $\mu$ and deviance for (a) the CA1 Rim, (b) CA1 Middle and (c) CA1 Core data from Upton et al. (2003) using Bootstrap method A with 5000 bootstraps. $\beta$ and $\mu$ do not show normal distributions for any of the data.

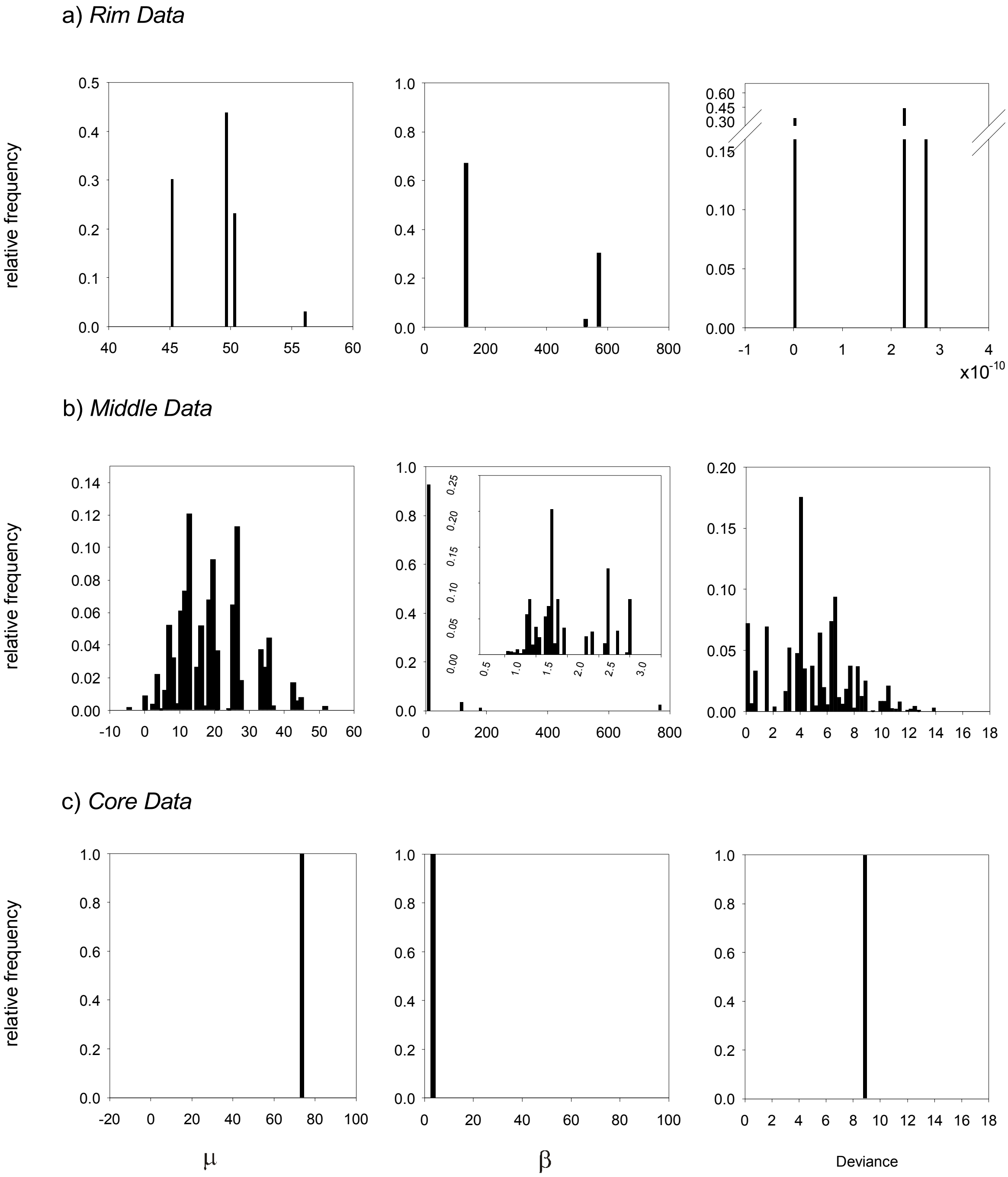


**Figure 4.** Bootstrapped empirical distributions of $\beta$, $\mu$ and deviance for (a) the CA1 Rim, (b) CA1 Middle and (c) CA1 Core data from Upton et al. (2003) using Bootstrap method B with 5000 bootstraps. Note the limited number of permutations, particularly for the Rim and Core data.

| | | | CA1 All | V209 |
|---|---|---|---|---|
| | | *n* | 64 | 986 |
| | | *m* | 17 | 18 |
| | | *μ* | 50 | 20.4 |
| | | *ß* | 2.85 | 9.39 |
| | | $^{G}D$** | 14.53 | 3.79 |
| *99% Confidence intervals* | *Upton* | *μ* | (28,71) | (17.5,23.3) |
| | | *ß* | (1.46,4.74) | (7.38,12.02) |
| *Bootstrap Method A* | *BCa* | *μ* | (27.6,72.5) | (17.6,23.0) |
| | | *ß* | (1.12,3.98) | (7.25,11.51) |
| | *Percentiles* | *μ* | (27.5,72.3) | (17.7,23.0) |
| | | *ß* | (1.71,5.54) | (7.73,12.36) |
| *Bootstrap Method B* | *BCa* | *μ* | (28.7,67.3) | (17.6,23) |
| | | *ß* | (1.38,3.92) | (7.16,11.5) |
| | *Percentiles* | *μ* | (29.2,67.8) | (17.6,23.0) |
| | | *ß* | (1.91,4.87) | (7.72,12.3) |

** $^{G}D$ is equivalent to Upton's "C".

**Table 3.** Results for the combined CA1 data in Upton et al. (2003) and for sample V209. 25000 bootstrap resamples were made in each case.

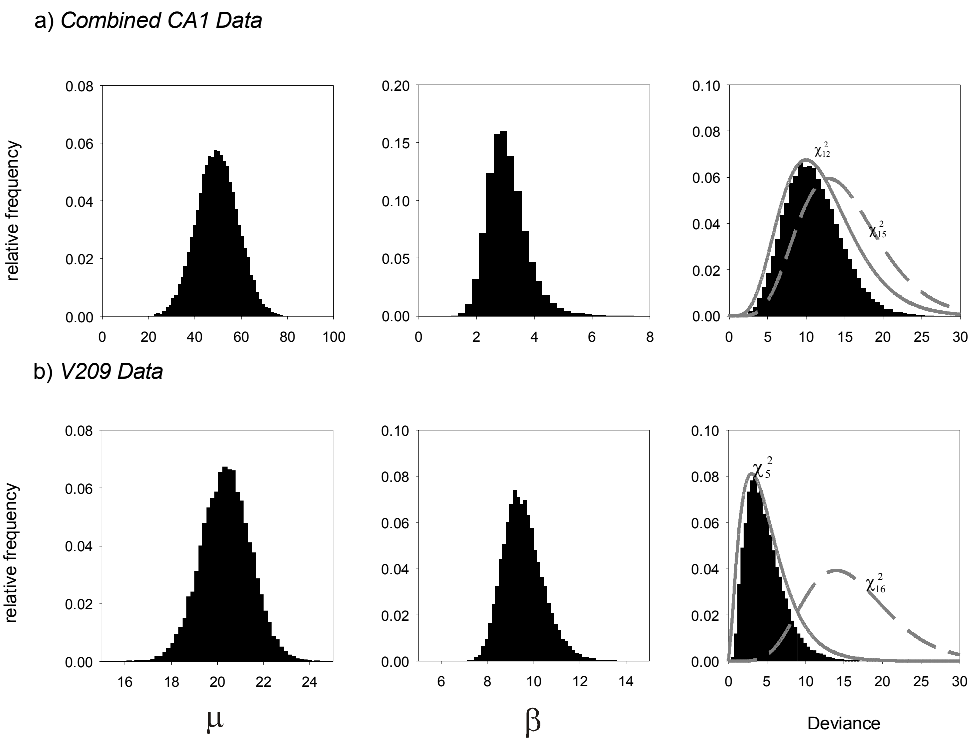


**Figure 5.** Bootstrapped empirical distributions of $\beta$, $\mu$ and deviance for (a) the combined CA1 data from Upton et al. (2003), and (b) sample V209 using Bootstrap method A with 25000 bootstraps. The distributions of $\beta$ and $\mu$ show normal and log-normal distributions respectively. For the deviance plots, the best fitting $\chi^2$ distributions are shown with solid lines and the $\chi^2$ distributions with *m*-2 degrees of freedom are shown with dashed lines.

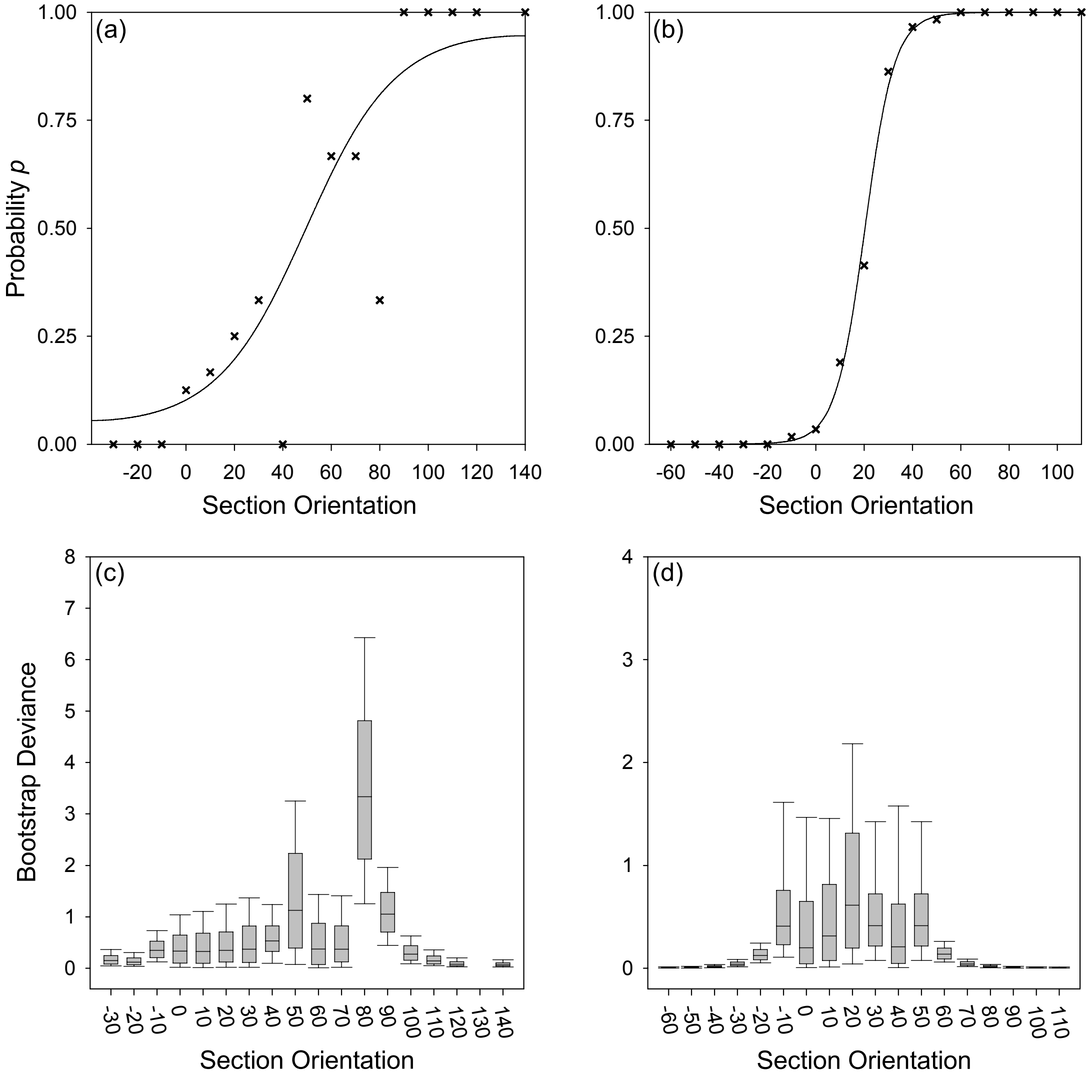


**Figure 6.** Probability plots showing the fit of the models to the observed probabilities (crosses) for (a) the combined CA1 data from Upton et al. (2003), and (b) V209 data. Box plots of bootstrap deviance are shown in (c) and (d). The V209 has fewer section orientations with significant deviance.

the correlations would be approximately zero. Furthermore, the variability for each orientation is non-uniform as orientations further away from the FIA will have very little or no variability as compared to those closest to the FIA, as shown in Figure 6e-f. These results suggest that the goodness-of-fit test of Upton et al. (2003) is flawed. They assume that there are two fixed parameters ($\beta$ and $\mu$) and that the probability of success can vary in each of *m* orientations; this is clearly not the case.

# 7 Conclusions

We examined methods for assessing the validity of logistic models for cyclic data, using bootstrapping methods. We have demonstrated that the goodness-of-fit test prescribed by Upton et al. (2003) is inappropriate because it does not adjust the degrees of freedom to take into account the dependencies within the model that contribute towards the model deviance.

In assessing the validity of the bootstrapped model estimates, we recommend examining the distributions of $\beta$ and $\mu$ as well as the deviance. Primarily, the distributions of $\beta$ and $\mu$ should be approximately normal, and, secondly the deviance distribution should approximate a $\chi^2$ type distribution. Further work to determining the null distribution of partially dependant deviance is still required to ascertain a formal test for the model goodness-of-fit. When determining the confidence intervals for $\beta$ and $\mu$, we recommend using the percentile method unless the distributions of $\beta$ and $\mu$ show departure from normality; in this case the $BC_a$ method should be used.

In employing bootstrapping techniques we encountered challenges using the sparse data used by Upton et al. (2003). In some circumstances the data was too sparse to assess model validity using the bootstrap. In others the resampling method for the bootstrap had a substantial influence in determining model validity. For our data we found that:

- For sparse data with $n_i = 1$, bootstrap on all observations (Method A);
- For sparse data with some 5~> $n_i > 1$ near the FIA, bootstrap within the observations (Method B);
- For rich data with $n_i$ ~> 4 for most orientations, bootstrap on all observations (Method A).

Caution must be applied where $n_i <$ ~5 and an indication of whether there is sufficient data can be gained by comparing the bootstrap distributions of $\beta$ and $\mu$, which should approximate a normal distribution.

| Deg | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 | 120 | 130 | 140 | 150 | 160 | 170 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1.00 | 0.17 | -0.03 | 0.05 | 0.00 | 0.05 | -0.04 | -0.04 | -0.05 | -0.05 | -0.05 | -0.05 | -0.05 | -0.06 | -0.06 | -0.06 | -0.06 | 0.09 |
| **10** | 0.17 | 1.00 | 0.47 | -0.08 | 0.02 | -0.08 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 | 0.04 | -0.16 |
| **20** | -0.03 | 0.47 | 1.00 | -0.07 | 0.06 | -0.07 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| **30** | 0.05 | -0.08 | -0.07 | 1.00 | 0.16 | 1.00 | 0.10 | 0.10 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.08 | 0.08 | 0.08 | 0.08 | -0.05 |
| **40** | 0.00 | 0.02 | 0.06 | 0.16 | 1.00 | 0.16 | -0.04 | -0.04 | -0.03 | -0.03 | -0.03 | -0.03 | -0.03 | -0.03 | -0.03 | -0.03 | -0.02 | 0.08 |
| **50** | 0.05 | -0.08 | -0.07 | 1.00 | 0.16 | 1.00 | 0.10 | 0.10 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.08 | 0.08 | 0.08 | 0.08 | -0.05 |
| **60** | -0.04 | 0.03 | 0.10 | 0.10 | -0.04 | 0.10 | 1.00 | 1.00 | 0.99 | 0.97 | 0.96 | 0.95 | 0.94 | 0.93 | 0.91 | 0.89 | 0.86 | 0.09 |
| **70** | -0.04 | 0.03 | 0.10 | 0.10 | -0.04 | 0.10 | 1.00 | 1.00 | 1.00 | 0.99 | 0.98 | 0.97 | 0.97 | 0.96 | 0.94 | 0.92 | 0.89 | 0.09 |
| **80** | -0.05 | 0.03 | 0.10 | 0.09 | -0.03 | 0.09 | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 0.98 | 0.97 | 0.96 | 0.94 | 0.91 | 0.09 |
| **90** | -0.05 | 0.03 | 0.10 | 0.09 | -0.03 | 0.09 | 0.97 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.98 | 0.97 | 0.96 | 0.93 | 0.09 |
| **100** | -0.05 | 0.03 | 0.10 | 0.09 | -0.03 | 0.09 | 0.96 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.98 | 0.97 | 0.94 | 0.09 |
| **110** | -0.05 | 0.03 | 0.10 | 0.09 | -0.03 | 0.09 | 0.95 | 0.97 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.97 | 0.95 | 0.09 |
| **120** | -0.05 | 0.03 | 0.10 | 0.09 | -0.03 | 0.09 | 0.94 | 0.97 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.98 | 0.96 | 0.09 |
| **130** | -0.06 | 0.04 | 0.11 | 0.08 | -0.03 | 0.08 | 0.93 | 0.96 | 0.97 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.97 | 0.09 |
| **140** | -0.06 | 0.04 | 0.11 | 0.08 | -0.03 | 0.08 | 0.91 | 0.94 | 0.96 | 0.97 | 0.98 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 0.10 |
| **150** | -0.06 | 0.04 | 0.11 | 0.08 | -0.03 | 0.08 | 0.89 | 0.92 | 0.94 | 0.96 | 0.97 | 0.97 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 0.10 |
| **160** | -0.06 | 0.04 | 0.11 | 0.08 | -0.02 | 0.08 | 0.86 | 0.89 | 0.91 | 0.93 | 0.94 | 0.95 | 0.96 | 0.97 | 0.99 | 1.00 | 1.00 | 0.10 |
| **170** | 0.09 | -0.16 | 0.11 | -0.05 | 0.08 | -0.05 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.10 | 0.10 | 0.10 | 1.00 |

**Table 4.** Deviance correlations between sampled orientations for the V209 data.

# References

Bell, T. H., Forde, A. & Wang, J. 1995. A new indicator of movement direction during orogenesis: measurement technique and application to the Alps. *Terra Nova* **V. 7**, 500-508.

Bell, T. H., Hickey, K. A. & Upton, G. J. G. 1998. Distinguishing and correlating multiple phases of metamorphism across a multiply deformed region using the axes of spiral, staircase and sigmoidal inclusion trails in garnet. *Journal of Metamorphic Geology* **16**(6), 767-794.

Dobson, A. 1990. *An Introduction to Generalized Linear Models*. Chapman & Hall, London.

Efron, B. & Tibshirani, R. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Hayward, N. 1990. Determination of early fold axis orientations in multiply deformed rocks using porphyroblast inclusion trails. *Tectonophysics* **V. 179**, 353-369.

Upton, G. J. G., Hickey, K. A. & Stallard, A. 2003. Regression models for cyclic data. *Royal Statistical Society Journal. Series C: Applied Statistics.* **52**(Part 2), 227-235.