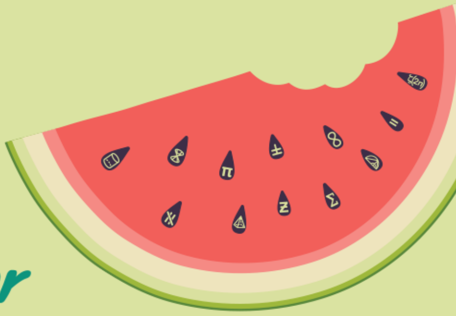
AMSI SUMMER RESEARCH SCHOLARSHIPS 2024–25

Get a taste for Research this Summer



Practical tools for assessing closed population capture–recapture models using the R-package DHARMa

Bernice Laitly

Supervised by Jakub Stoklosa

The University of New South Wales

Jane Street

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

ABSTRACT

1. Capture–recapture models are crucial for estimating demographic parameters when detection is imperfect.
2. Assessing model fit is essential for valid statistical inference, but the non-standard distributions commonly encountered in capture–recapture models often render standard software inadequate, requiring custom programming.
3. To address this, we employ a partial likelihood approach conditioned on initial capture, transforming the problem into a binomial model. This enables goodness-of-fit testing for closed capture–recapture models using the `R` package `DHARMa` (Diagnostics for HierArchical Regression Models).
4. We provide practical guidance on implementing this approach and demonstrate `DHARMa`'s application using both simulated and real-world capture–recapture data.
5. Our findings show that `DHARMa` effectively detects misspecified heterogeneity in capture probabilities within closed capture–recapture models. Additionally, under the partial likelihood framework, the power to detect goodness-of-fit issues improves with an increasing number of capture occasions.

1 INTRODUCTION

Capture–recapture (CR) models are fundamental to population ecology, offering a means of estimating demographic parameters when individuals are not always detected (McCrea & Morgan, 2014). These methods involve tracking marked individuals over time through repeated captures or sightings, resulting in capture histories and associated covariates. It is crucial to recognize that the observed data are inherently conditional on individuals being captured at least once. When populations are closed (no births, deaths, immigration, or emigration) during the study period, a key objective is often to estimate the unknown population size.

Otis *et al.* (1978) pioneered a suite of closed population CR models that accounted for individual heterogeneity, time effects, and behavioral responses. Huggins (1989) later proposed a more general conditional likelihood approach, offering greater flexibility by avoiding restrictive distributional assumptions and addressing potential identifiability problems. Existing methodologies for fitting closed population CR models include sample coverage (Lee & Chao, 1994), finite mixture models (Pledger, 2000; Dorazio & Royle, 2003; Pledger, 2005; Morgan & Ridout, 2008), empirical likelihood (Liu *et al.*, 2017) and many others. In this paper, we concentrate on the Huggins (1989)'s heterogeneity-only model (M_h), where capture probabilities are modeled as functions of unknown parameters and individual covariates, which we estimate using maximum likelihood or Bayesian methods.

Regardless of the chosen statistical framework, assessing model fit is critical for drawing valid demographic inferences from CR data (Gimenez *et al.*, 2018). Early approaches to formal goodness-of-fit (GOF) testing for closed CR models were often ad hoc and relied on chi-squared tests comparing observed and expected capture frequencies (e.g., see Carothers, 1971; Otis *et al.*, 1978). However, these methods often suffered limitations due to the sparse nature of CR data, particularly with small or highly heterogeneous populations. The inherent dependence among captures further complicates the application of standard statistical tests. Furthermore, the variety of model structures (e.g., accounting for heterogeneity via observed covariates or latent variables) necessitates tailored GOF procedures.

The R-package DHARMa (Diagnostics for Hierarchical Regression Models, Hartig, 2024) offers a useful set of GOF tests and readily interpretable residual plots for a wide range of models, including generalized linear mixed models. Residual plots have the potential as a diagnostic tool for CR models, however, DHARMa's current capabilities do not extend to the positive binomial distribution, which is fundamental to the Huggins (1989)'s conditional likelihood approach. This limitation often necessitates custom programming, as illustrated by Warton *et al.* (2017), who developed bespoke code for graphical diagnostics in occupancy-detection models.

To address this issue, we employ the partial likelihood methodology described by Stoklosa *et al.* (2011). This approach, which conditions on the initial capture event, results in a binomial model amenable to parameter estimation based on recapture data. This formulation facilitates the use of standard software packages such as DHARMa for GOF model evaluation, as DHARMa is specifically designed for residual analysis in binomial models, among others.

After outlining the methodological background, we illustrate DHARMa's utility using simulated datasets with different levels of individual heterogeneity, followed by analyses of two real CR datasets. Our results demonstrate that DHARMa effectively detects lack of fit in CR models, especially when heterogeneity in capture probabilities is incorrectly specified. The reliability of these diagnostics improves as the number of capture occasions increases, suggesting that studies with more sampling periods will benefit most from this approach.

BL conducted the computer programming, data analysis, writing, and checking earlier drafts of the report.

2 METHODS

2.1 Notation and Models

We consider a closed population of unknown size N and a CR experiment with τ capture occasions ($j = 1, \ldots, \tau$). Assuming independence among individuals, we define two random variables for the ith individual: Y_i, the number of captures, and t_i, the time of the first capture.

We model heterogeneity in capture probabilities using individual-specific covariates X_i, assumed to be independent and identically distributed. Let $\boldsymbol{X}_i$ be the covariate vector for individual i and $\boldsymbol{\beta}$ the associated parameter vector. The number of captures for individual i is then $Y_i \sim \text{Bin}(\tau, p_i(\boldsymbol{\beta}))$, where $p_i(\boldsymbol{\beta}) = H(\boldsymbol{\beta}^T \boldsymbol{X}_i)$ and $H(u) = \exp(u)/(1 + \exp(u))$ is the logistic function. The probability of individual i being captured at least once is $\pi_i(\boldsymbol{\beta}) = 1 - \{1 - p_i(\boldsymbol{\beta})\}^{\tau}$. Captured individuals are indexed by $i = 1, \ldots, D$, and uncaptured individuals by $i = D + 1, \ldots, N$ where D is the number of distinct captures in the study. As in Stoklosa *et al.* (2011), the full likelihood (proportional to) is given by

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{p_i(\boldsymbol{\beta})^{Y_i}\{1 - p_i(\boldsymbol{\beta})\}^{\tau - Y_i}}{\pi_i(\boldsymbol{\beta})} \prod_{i=1}^{D} \pi_i(\boldsymbol{\beta}) \prod_{i=D+1}^{N} \{1 - \pi_i(\boldsymbol{\beta})\}.$$

The first product represents the contribution of captured individuals, the second the probability of being captured at least once, and the third the probability of not being captured. Because the covariates of uncaptured individuals are unknown, this full likelihood is not directly calculable. Huggins (1989)'s conditional likelihood, which conditions on the event that the first D individuals are captured, uses only the first product. It can be expressed as

$$\prod_{i=1}^{D} p_i(\boldsymbol{\beta})^{Y_i-1}\{1-p_i(\boldsymbol{\beta})\}^{\tau-t_i-(Y_i-1)} \prod_{i=1}^{D}\left[\frac{\{1-p_i(\boldsymbol{\beta})\}^{t_i-1}p_i(\boldsymbol{\beta})}{\pi_i(\boldsymbol{\beta})}\right]. \tag{1}$$

The first term in (1) constitutes the partial likelihood (Stoklosa *et al.*, 2011) which represents the probability of the observed recaptures, given the time of first capture. Specifically, conditional on t_i, the number of recaptures for individual i, $(Y_i - 1)$, follows a binomial distribution:

$$(Y_i - 1) \mid t_i \sim \text{Bin}(\tau - t_i, p_i(\boldsymbol{\beta})). \tag{2}$$

After obtaining $\hat{\boldsymbol{\beta}}$ from fitting model (2), the Horvitz–Thompson estimator: $\widehat{N} = \sum_{i=1}^{D} 1/\pi_i(\hat{\boldsymbol{\beta}})$ is often used to estimate population size. While Stoklosa *et al.* (2011) observed a potential minor efficiency loss in parameter estimation with partial likelihood, their simulations demonstrated substantial improvements with increasing τ. Furthermore, the practical advantages of leveraging existing software to fit complex models like generalized additive models (GAMs) or generalized linear mixed models (GLMMs) often outweigh efficiency concerns. Consequently, this work concentrates on assessing model fit using readily available software. Although parameter estimation can follow once a suitable model is identified, it is not the primary objective here.

2.2 Goodness-of-Fit assessment using `DHARMa`

The `R`-package `DHARMa` provides tools for assessing GOF for a variety of models, including, generalized linear models (GLMs), GLMMs, GAMs, etc., and can handle Bayesian software. It uses simulated (quantile) residuals to generate readily interpretable diagnostic plots and perform formal GOF tests. For technical and theoretical details on quantile residuals for regression models, see Dunn & Smyth (1996). By comparing the distribution of simulated residuals to the expected distribution under the null hypothesis (that the model adequately fits the data), analysts can use `DHARMa`'s functions to evaluate model adequacy. Rather than providing a single "yes/no" answer, `DHARMa` offers a comprehensive suite of diagnostic tools to assess model fit thoroughly.

Several graphical diagnostics are available in `DHARMa` to visually assess the distribution of simulated residuals. By default, these include: a quantile–quantile (QQ) plot, which compares the distribution of simulated residuals to a uniform distribution (deviations from the diagonal line indicate potential issues), and a plot of residuals vs. fitted values, used to check for patterns in residuals across the range of predicted values. Plots of residuals vs. predictor variables, which examine relationships between residuals and specific predictors can also be obtained. Patterns in these latter two plots also suggest departures from model assumptions. Other diagnostics plots are also available but they are not explored here.

`DHARMa` also includes statistical tests to assess the distribution of simulated residuals formally. For example, the Kolmogorov–Smirnov (KS) test checks if the residuals are significantly different from a uniform distribution. Beyond uniformity testing, `DHARMa` assists in diagnosing and handling overdispersion/underdispersion (Dispersion test), a common issue in many models, and in detecting residual outliers (Outlier test). Other tests also include zero-inflation, residual spatial, temporal, and phylogenetic autocorrelation but they are not considered here.

2.3 Guidelines for practitioners

Prior to model fitting, we recommend evaluating the assumption of population closure using the test proposed by Stanley & Burnham (1999). This can be easily implemented within the secr package (Efford, 2024). However, as emphasized by Efford (2024), the Stanley and Burnham test should be interpreted cautiously. It can suggest non-closure in truly closed populations if, for example, trap response is present. Furthermore, the test is sensitive to individual heterogeneity, a factor that should prompt consideration of the species' biology.

If a closed population is considered, then any binomial model of interest that is compatible with DHARMa can be fitted to the closed population CR data discussed in Section 2.1, and its GOF evaluated. It's important to note however that using $Y_{i=1,...,D}$ directly in a binomial model will lead to poor fit because $Y_{i=1,...,D}$ follows a positive binomial distribution, not a standard binomial.

The following R-code demonstrates how to construct recapture data (Y representing the number of captures, t1 the time of first capture, X a covariate vector, and tau the number of capture occasions), fit a GLM with a covariate, and assess its GOF using DHARMa:

```r
recap_opp <- tau - t1[!t1 == tau]
Y_recap <- Y[!t1 == tau] - 1
X_recap <- X[!t1 == tau]

model <- glm(Y_recap / recap_opp ~ X_recap,
family = binomial(link = "logit"),
weights = recap_opp)

res0 <- simulateResiduals(fittedModel = model)
plot(res0)
```

Notice that in the code above, to use Model (2), individuals captured only on the last occasion are excluded to allow the calculation of recapture opportunities.

Figure 1 presents an example of DHARMa diagnostics applied to real CR data following the fit of a GLM. Further details on the study, data, and GOF analysis can be found in Section 3.2.2. Small p-values (highlighted in red) on the QQ-plot, as well as red-highlighted curves in the residuals vs. fitted values plot, signal potential model misfit. When assessing M_h-type CR models, all three diagnostic tests should be considered. In particular, the Dispersion test can detect overdispersion, which may indicate unmodeled individual heterogeneity in capture probabilities. Additionally, the residuals vs. fitted values plot should be carefully examined, as patterns in the residuals may reveal misspecification in the relationship between capture probabilities and covariates.

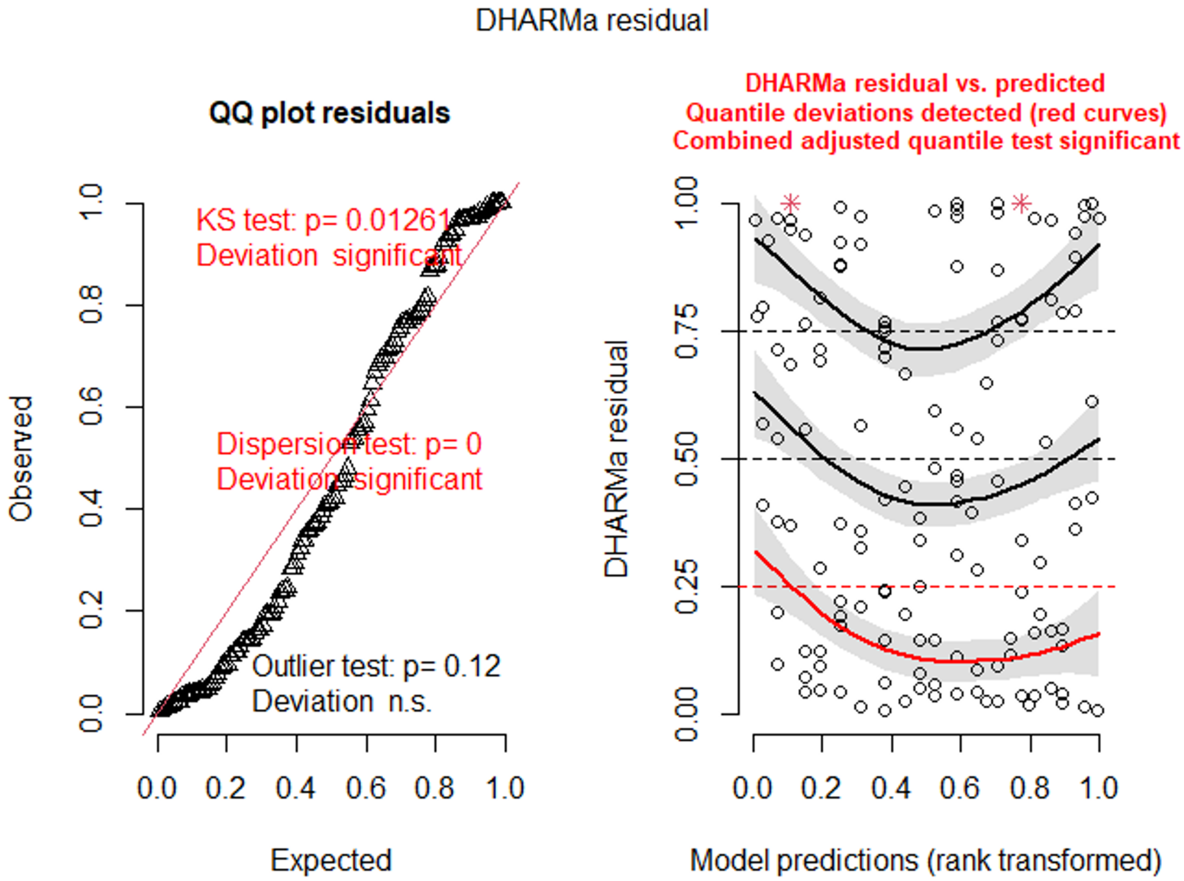


Figure 1: DHARMa *diagnostics when fitting a GLM to the Harvest mouse capture–recapture data (see Section 3.2.2 for further details). The figure gives a QQ plot (left panel), which compares the distribution of simulated residuals to a uniform distribution (deviations from the diagonal line indicate potential issues) and a plot of residuals vs. fitted values (right panel), used to check for patterns in residuals across the range of predicted values. A Kolmogorov–Smirnov (KS) test to check if residuals are significantly different from a uniform distribution, an overdispersion/underdispersion test (Dispersion test), a common issue in many models, and an Outlier test in detecting residual outliers.*

3 APPLICATION TO DATA

How effective are DHARMa diagnostic tools at detecting a lack of fit for CR data? This question was explored, using simulations and two-real CR data sets.

3.1 Simulations

First, we evaluated DHARMa's performance in detecting GOF issues in various simulation scenarios/cases where heterogeneity was present in capture probabilities.

Under a closed population setting, we generated CR data for three scenarios where the true model for p_i had the following structures:

- **Case 1 (two covariates):** $p_i = H(\beta_0 + \beta_1 X_i + \beta_2 W_i)$ where $W_i \sim N(1, 2)$ or $W_i \sim \text{Bern}(0.5)$. This scenario models capture probabilities influenced by two observed sources of individual heterogeneity.
- **Case 2 (random effect):** $p_i = H(\beta_0 + \beta_1 X_i + Z_i)$ where $Z_i \sim N(0, 1)$. This scenario incorporates both an observed (X_i) and an unobserved (Z_i) source of individual heterogeneity.
- **Case 3 (non-linearity):** $p_i = H(\beta_0 + \beta_1 \sin(X_i))$. This scenario explores a non-linear, smooth relationship between capture probability and the observed covariate X_i.

In each case above, we generated data from a population of size $N = 500$, with an individual continuous covariate X_i drawn from a standard normal distribution, $X_i \sim N(0, 1)$. Capture occasions were set at either $\tau = 7$ or $\tau = 15$. The true model parameters $(\beta_0, \beta_1, \beta_2)$ varied depending on the specific simulation scenario. For each scenario and parameter/covariate combination, we generated 20 datasets. We then fit both the correctly specified model (as described above) and a misspecified model that omitted a key component (descriptions given below). This misspecification allowed us to evaluate the performance of the `DHARMa` package in detecting a lack of fit. Specifically, we counted the number of *any* significant issues reported by `DHARMa`'s three diagnostic tests for each model fit. We also reported the average number of unique individuals detected ($\bar{D}$) to quantify the average number of individuals captured across the simulations.

We also considered a fourth scenario, Case 4 (open population), where data were generated under a true open population design. In this scenario, individuals could permanently leave the population with a survival rate ϕ, and new individuals could enter at a rate ψ on each capture occasion. Capture probabilities were modeled as $p_i(\boldsymbol{\beta}) = H(\beta_0 + \beta_1 X_i)$ where $X_i \sim N(0, 1)$ or a Bernoulli distribution with a probability of success of 0.5, $X_i \sim \text{Bern}(0.5)$. Here, we set $N = 200$ and $\tau = 12$, and $\phi = \psi = 0.8$ to maintain a balanced influx and efflux of individuals. This simulation was designed to assess the robustness of `DHARMa` when the population structure is misspecified (i.e., when a closed population model is fit to open population data).

The complete simulation results, encompassing the number of GOF issues/outcomes determined by `DHARMa` for each simulation case and parameter combination, are documented in Appendix A, presented in tabular format as Tables A1–A10.

3.1.1 Simulation case 1 (two covariates)

We fitted standard GLMs using (2). Both the correctly specified (Model 1) and misspecified (Model 2) models included X_i as a linear term, but the misspecified model omitted W_i. Figure 2 presents bar charts comparing the number of GOF issues reported by `DHARMa` for both model fits. The top panel displays results for $W_i \sim N(1, 2)$, while the bottom panel shows results for $W_i \sim Bern(0.5)$. The left panel corresponds to $\tau = 7$ capture occasions and the right panel to $\tau = 15$.

Omitting the covariate W_i in the GLM fit (Model 2) resulted in frequent diagnostic flags from `DHARMa`, with the dispersion test consistently indicating model misspecification. This issue was less pronounced when W_i was binary, although misspecification was still detected in approximately half of the simulations. Other tests and quantile deviations also more frequently identified significant issues in the misspecified model compared to the correctly specified model. Furthermore, the performance of `DHARMa` improved with increasing τ for both models, suggesting greater reliability for a large number of capture occasions.

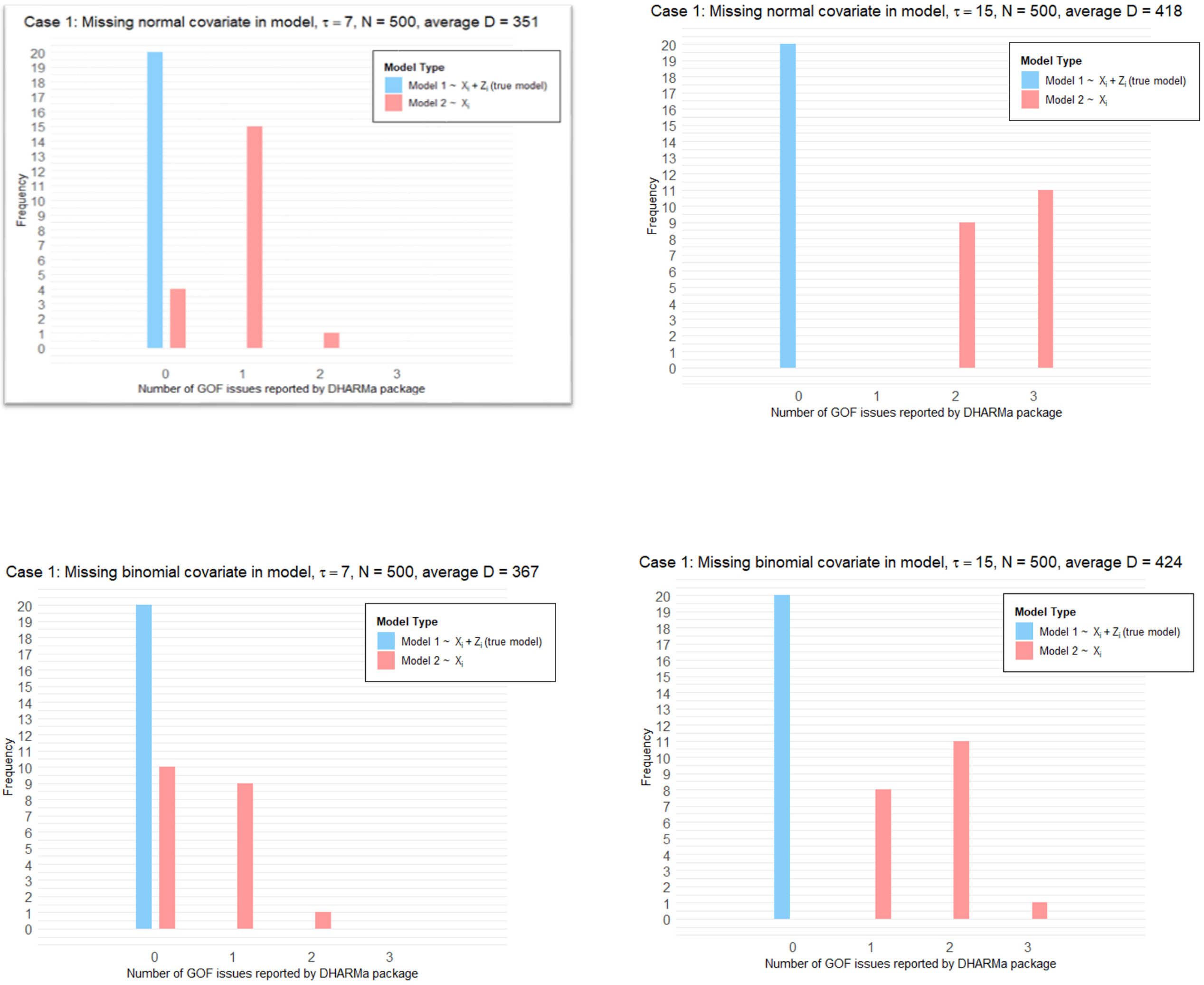


Figure 2: *(Simulation case 1) Bar charts comparing the number of goodness-of-fit (GOF) issues reported by* `DHARMa` *for the correct model (Model 1) and misspecified model (Model 2). The top panel displays results for the case where* $W_i \sim N(1, 2)$*, while the bottom panel shows results for* $W_i \sim \text{Bern}(0.5)$*. The left panel corresponds to* $\tau = 7$ *capture occasions and the right panel to* $\tau = 15$*.*

3.1.2 Simulation case 2 (random effect)

Next, we fitted both a GLM and a GLMM using (2) for Case 2. The misspecified GLM (Model 2) omitted the random effect present in the true model. The GLMM (Model 1) was fit using the `glmer()` function from the `lme4` package. Figure 3 presents bar charts comparing the number of GOF issues reported by `DHARMa` for both the correctly specified GLMM and the misspecified GLM.

When including a random effect, Model 1, `DHARMa` reported minimal diagnostic flags across both $\tau = 7$ and $\tau = 15$, whereas excluding the random effect (Model 1) there were frequent issues. These issues were more prominent for $\tau = 15$. Occasionally, when the dispersion test was deemed significant for the correct model, the actual shape of the QQ plot still indicated that the model gave a good fit, as compared to the incorrect model where the QQ plot shape also supports the significant test, see Figure 4.

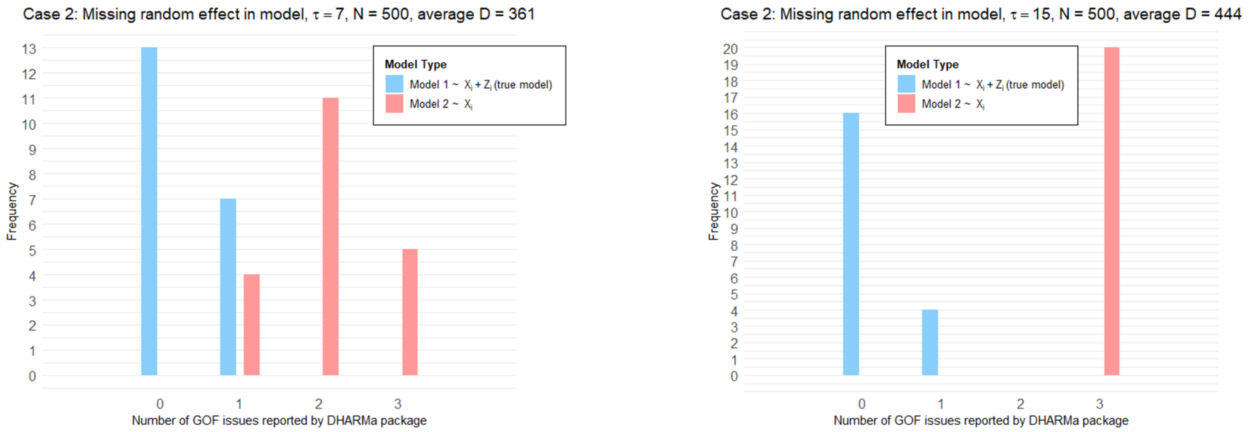


Figure 3: *(Simulation case 2) Bar charts comparing the number of goodness-of-fit (GOF) issues reported by* DHARMa *when fitting the correct (GLMM, Model 1) and incorrect model (GLM, Model 2). The left panel corresponds to $\tau = 7$ capture occasions and the right panel to $\tau = 15$.*

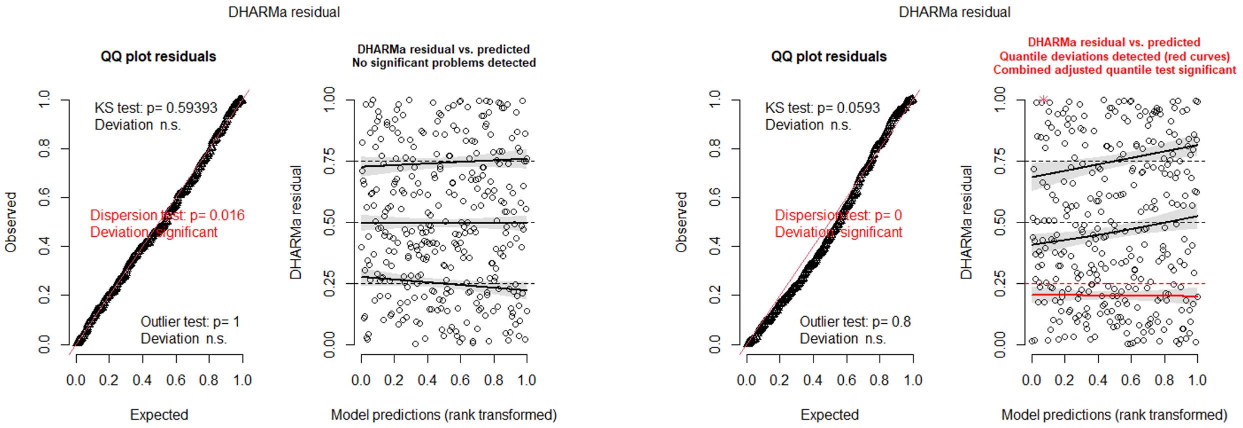


Figure 4: *(Simulation case 2) Diagnostic plots for a selected simulation when fitting the correct model (left panel) and the wrong model (right panel). Notice the difference in the shape of the QQ plot even though both have significant dispersion tests.*

When the random effect was correctly included, DHARMa reported minimal diagnostic warnings at both $\tau = 7$ and $\tau = 15$. Conversely, omitting the random effect resulted in frequent warnings, particularly at $\tau = 15$. Occasionally, although the dispersion test flagged the correctly specified model as significant, the corresponding QQ plot indicated a good fit. In contrast, for the misspecified model, the QQ plot visually confirmed the lack of fit indicated by the significant test, as illustrated in Figure 4.

3.1.3 Simulation case 3 (non-linearity)

We fitted both a GLM and a GAM using (2) for Case 3. The misspecified GLM (Model 2) failed to account for the non-linear relationship between capture probability and X_i. The GAM (Model 1) was fit using the `gam()` function from the `mgcv` package. Figure 5 presents bar charts comparing the number of GOF issues reported by DHARMa for both the correctly specified GAM and the misspecified GLM.

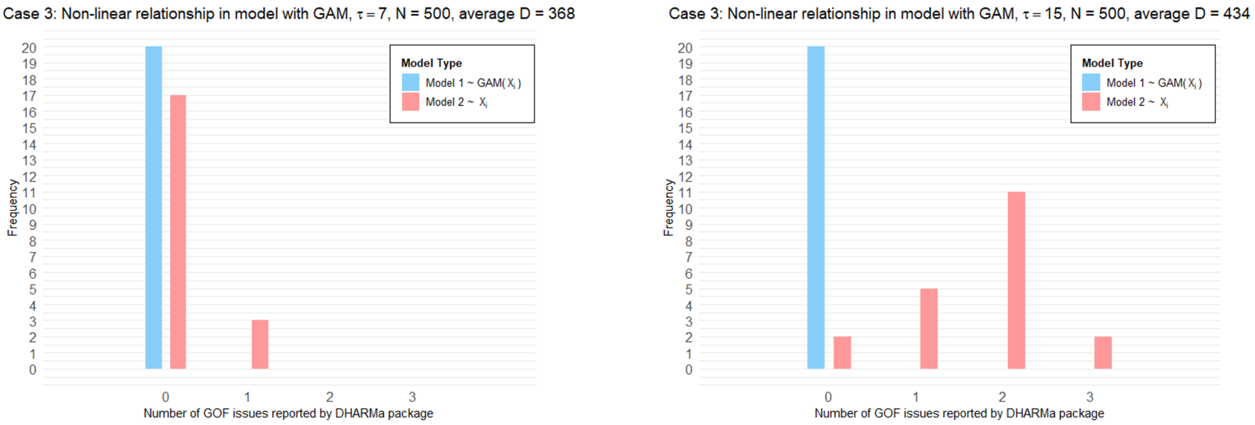


Figure 5: *(Simulation case 3) Bar charts for the number of GOF issues reported by* `DHARMa` *when fitting the correct (GAM, Model 1) and incorrect model (GLM, Model 2). The left panel is for* $\tau = 7$ *and the right panel is for* $\tau = 15$.

The GAM performed well, with `DHARMa` reporting no diagnostic warnings in either scenario. In contrast, the misspecified GLM performed poorly, with `DHARMa` consistently flagging significant tests and quantile deviations.

3.1.4 Simulation case 4 (open population)

Finally, we fit two misspecified models, a GLMM (Model 1) and a GLM (Model 2), using (2), both incorrectly assuming a closed population structure. For the GLM, `DHARMa` consistently flagged issues with the dispersion, outlier, and KS tests. These flags indicated substantial deviations from the expected distribution, demonstrating the poor performance of closed-population GLMs when applied to open population data.

When fitting the GLMM, however, `DHARMa` largely failed to detect significant GOF issues, despite the model's incorrect specification. Although residual plots and diagnostic tests (e.g., QQ plots and KS tests) suggested an adequate fit, this result highlights a potential limitation of `DHARMa` in identifying structural misspecification when using `glmer()` on open population data. Figure 6 provides diagnostic plots from a selected representative simulation, comparing the GLM (left panel) and GLMM (right panel) fits. While `DHARMa` successfully identified issues with the GLM, it failed to detect the same underlying structural problems when individual random effects were included in the GLMM. This underscores a potential weakness of `DHARMa` in assessing/evaluating GOF for open population models.

3.2 Real Data Examples

To assess `DHARMa`'s ability to detect GOF issues in real-world scenarios, we analyzed two well-known CR datasets previously studied in the literature. The first dataset originates from a study of taxi cabs in Edinburgh, Scotland (Carothers, 1973), and the second from a CR experiment on harvest mice in Taiwan (Stoklosa *et al.*, 2011).

3.2.1 Example 1: Taxi cab data

The taxi cab dataset comprised capture histories from $\tau = 10$ capture occasions. A "capture" was recorded each time a taxi cab passed a pre-defined location in Edinburgh. The population was assumed closed (supported by the Stanley

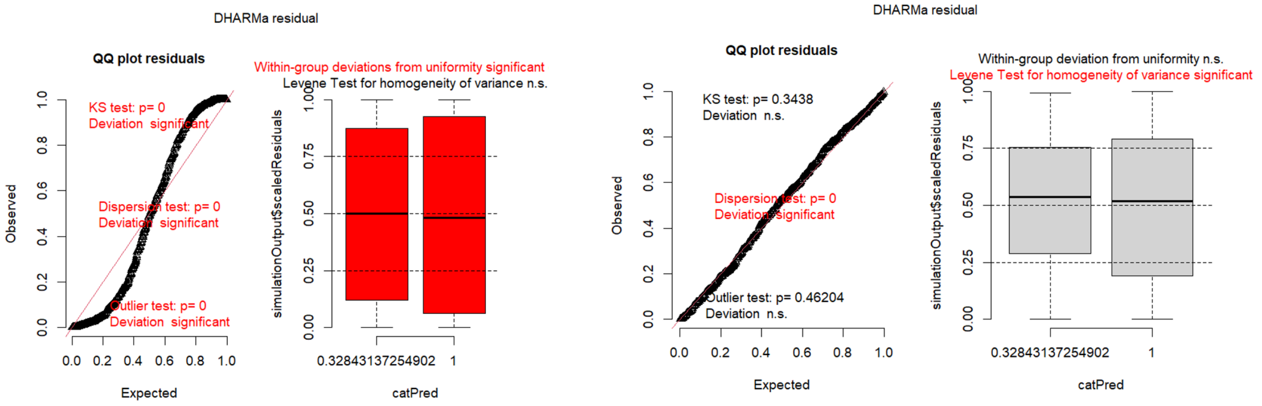


Figure 6: *(Simulation case 4) Diagnostic plots for a selected simulation when fitting a GLM (left panel) and a GLMM (right panel). The GLMM resulted in a significant dispersion test; however, the QQ plot does not show strong evidence of misfit, similar to the correctly specified models examined earlier. This suggests that the diagnostic plots may not fully reflect this issue despite the model being incorrect due to the data originating from an open population.*

and Burnham closure test), with a known size of $N = 420$. Given that only capture histories were available (i.e., no individual covariates were measured), an intercept-only GLM with binomial logistic regression was used to model capture probabilities. Figure 7 displays the resulting `DHARMa` diagnostic plots.

`DHARMa` detected no significant goodness-of-fit issues. The QQ plot showed no substantial deviations, and the KS test, overdispersion and underdispersion test, and outlier test all yielded non-significant p-values. The residual plot confirmed an adequate model fit, with no within-group deviations from uniformity. Overall, the model provided a reasonable fit to the taxi cab data, with `DHARMa` raising no diagnostic warnings.

3.2.2 Example 2: Harvest mouse data

The harvest mice dataset contained capture histories and individual covariates conducted across $\tau = 14$ capture occasions. Two key covariates were available: body weight (standardized weight at first capture) and sex (a binary variable: 1 for males, 0 for females). After conducting the Stanley and Burnham closure test, initial analysis strongly suggested an open population structure (p-value < 0.001), although the high heterogeneity in the data introduces some uncertainty in this result (Stoklosa *et al.*, 2011).

We fit three models to the data (each model includes the sex covariate): a GLM, a GLMM with a random effect and a linear term for body weight, and a GLMM with a random effect and a quadratic (squared) term for body weight. The quadratic term was included based on AIC, which indicated improved model fit. In Figures 1 and 8, we plot the `DHARMa` diagnostics for each model.

While the initial GLM and GLMM with the linear body weight term showed some lack of fit (explain what specifically was wrong with the linear models), the GLMM with the quadratic body weight term showed substantial improvement, with no goodness-of-fit issues detected by `DHARMa`. A curved downward pattern observed in the `DHARMa` residual vs. predicted plot (right panel) for the linear models was addressed by including the squared weight term, which further improved the model fit and straightened the residual plot. The final model (GLMM with the quadratic term) was selected

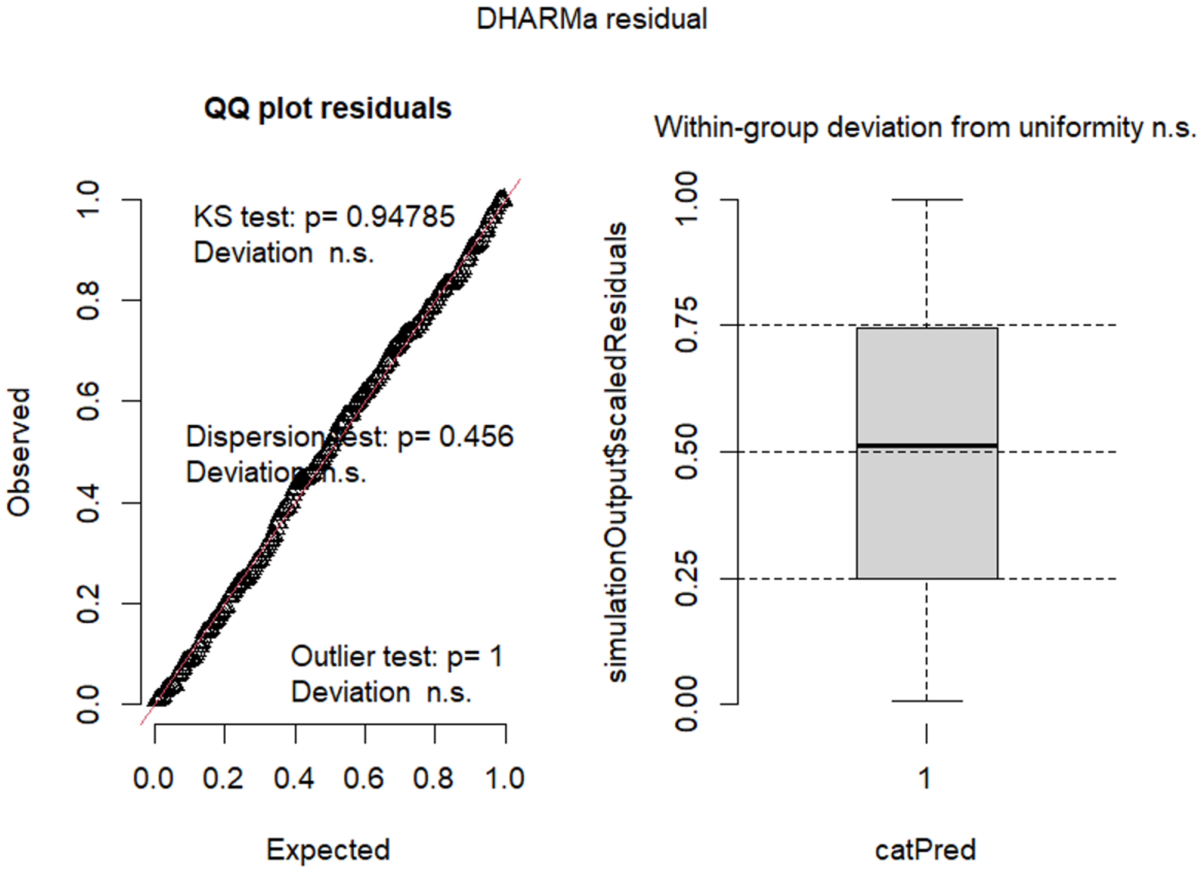


Figure 7: `DHARMa` *diagnostics when fitting a GLM to the taxi cab CR data.*

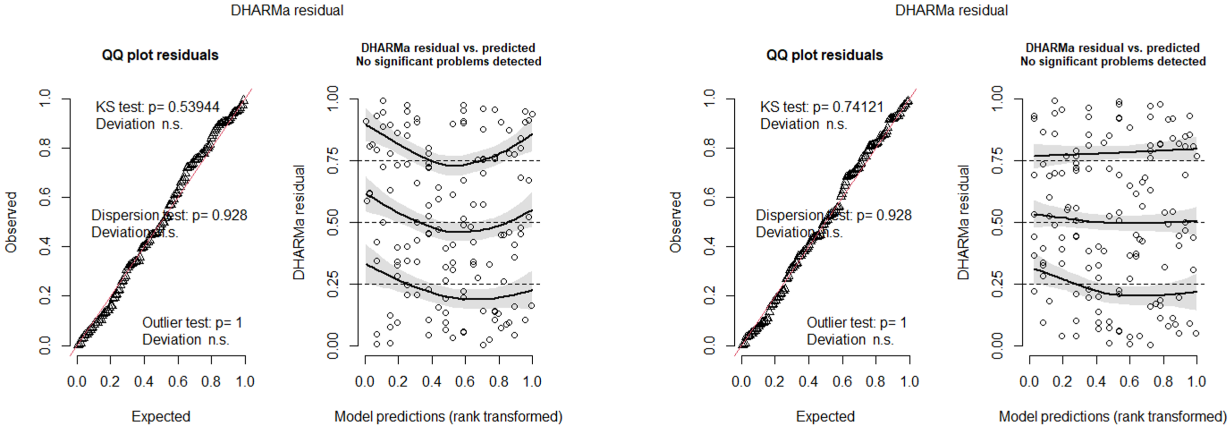


Figure 8: `DHARMa` *diagnostics plots when fitting a GLMM with a random effect and a linear term for body weight (left panel), and a GLMM with a random effect and a quadratic (or squared) term for body weight (right panel) to the Harvest mouse data (both models also includes the sex covariate).*

as the best fit based on `DHARMa` diagnostics and AIC comparisons.

4 DISCUSSION

Our study addresses a gap in GOF assessment/diagnostics for closed population CR models. Using a partial likelihood approach with recapture data, we leverage specialized software to provide improved GOF evaluation. While diverse analysis methods exist for closed populations (see Section 1), they typically lack the ease and flexibility, particularly for visual diagnostics, that our approach offers. Interpreting graphical diagnostics can however be subjective, and these methods are often used in conjunction with other GOF tests. While `DHARMa` serves as a powerful diagnostic tool for identifying potential GOF issues, it does not guarantee that the selected model is correct or optimal, nor does every flagged issue necessarily indicate a model deficiency.

In this study, we primarily focused on GOF assessment for closed population models (specifically the heterogeneity model, M_h). For closed populations, the `VGAM` package can fit all eight models described by Huggins (1989), including the most general model, M_{tbh}, and it provides residuals. However, `VGAM` model objects are currently incompatible with the `DHARMa` package. Further work is also needed to assess GOF when simpler models (e.g., M_h) are fitted to data generated from more complex models (e.g., with time and behavioral effects). We intend to address these limitations in future research.

For open populations models, Gimenez *et al.* (2018) introduced the `R`-package `R2ucare` which implements and expands upon existing tests based on contingency tables (Pollock *et al.*, 1985), providing a comprehensive tool for assessing the fit of the well-known Cormack-Jolly-Seber and other related open population models (Pradel *et al.*, 2003). It tests for the presence of transients, trap dependence, and overdispersion. To further enhance its utility, the addition of residual plots as a graphical diagnostic, similar to those provided by the `DHARMa` would be highly beneficial for model evaluation in open population studies.

Finally, we note that GOF assessment differs from model selection, though the two are related (Warton *et al.*, 2017). Model selection involves comparing the relative fit of candidate models (e.g., choosing the model with the lowest AIC). GOF, conversely, seeks an absolute measure: does the model adequately explain the observed data? For example, while model selection using AIC identifies the best relative fit, GOF asks whether the AIC value for that model is sufficiently low to suggest plausibility, perhaps by comparing it to a null distribution.

Acknowledgements

JS is supported by the Australian Research Council Discovery Project scheme DP210101923 and BL was supported by a 2024/25 AMSI Summer Research Scholarship.

References

Carothers, A. D. (1971). An examination and extension of Leslie's test of equal catchability. *Biometrics*, **27**, 615–630.

Carothers, A. D. (1973). The effects of unequal catchability on Jolly-Seber estimates. *Biometrics*, **29**, 79–100.

Dorazio, R. M. & Royle, J. A. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, **59**, 351–364.

Dunn, P. K. & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Efford, M. (2024). `secr`*: Spatially explicit capture-recapture models*. `R` package version 5.1.0.

Gimenez, O., Lebreton, J.-D., Choquet, R. & Pradel, R. (2018). R2ucare: An `R` package to perform goodness-of-fit tests for capture–recapture models. *Methods in Ecology and Evolution*, **9**, 1749–1754.

Hartig, F. (2024). `DHARMa`*: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. `R` package version 0.4.7.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika*, **76**, 133–140.

Lee, S.-M. & Chao, A. (1994). Estimating population size via sample coverage for closed capture–recapture models. *Biometrics*, **50**, 88–97.

Liu, Y., Li, P. & Qin, J. (2017). Maximum empirical likelihood estimation for abundance in a closed population from capture-recapture data. *Biometrika*, **104**, 527–543.

McCrea, R. S. & Morgan, B. J. T. (2014). *Analysis of Capture–Recapture Data.*. Chapman & Hall/CRC, London.

Morgan, B. J. T. & Ridout, M. S. (2008). A new mixture model for capture heterogeneity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**, 433–446.

Otis, D. L., Burnham, K. P., White, G. C. & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, **62**, 3–135.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, **56**, 434–442.

Pledger, S. (2005). The performance of mixture models in heterogeneous closed population capture–recapture. *Biometrics*, **61**, 868–876.

Pollock, K. H., Hines, J. E. & Nichols, J. D. (1985). Goodness-of-fit tests for open capture–recapture models. *Biometrics*, **41**, 399–410.

Pradel, R., Wintrebert, C. M. A. & Gimenez, O. (2003). A proposal for a goodness-of-fit test to the Arnason–Schwarz multisite capture–recapture model. *Biometrics*, **59**, 43–53.

Stanley, T. & Burnham, K. (1999). A closure test for time-specific capture-recapture data. *Environmental and Ecological Statistics*, **6**, 197–209.

Stoklosa, J., Hwang, W.-H., Wu, S.-H. & Huggins, R. M. (2011). Heterogeneous capture-recapture models with covariates: A partial likelihood approach for closed populations. *Biometrics*, **67**, 1659–1665.

Warton, D. I., Stoklosa, J., Guillera-Arroita, G., MacKenzie, D. I. & Welsh, A. H. (2017). Graphical diagnostics for occupancy models with imperfect detection. *Methods in Ecology and Evolution*, **8**, 408–419.

Appendix A: Results tables for all simulation studies

Table A1: *(Simulation case 1) Tabulated results comparing the number of goodness-of-fit (GOF) issues/outcomes reported by* `DHARMa` *for the correct model (Model 1, left column) and misspecified model (Model 2, right column) where* $N = 500, \tau = 7, W_i \sim N(1, 2)$ *and* $\bar{D} = 350.50$ *for 20 simulated data sets.*

	Outcome reported by `DHARMa`	
Run #	Model 1 (correct model, includes W_i)	Model 2 (misspecified model, excludes W_i)
1	None	KS test, dispersion test; 0.25
2	None	Dispersion test
3	None	Dispersion test
4	None	Dispersion test
5	None	Dispersion test
6	None	Dispersion test; 0.25
7	None	Dispersion test
8	None	Dispersion test; 0.25
9	None; 0.75	Dispersion test
10	None	None
11	None	Dispersion test
12	None	Dispersion test
13	None	Dispersion test
14	None; 0.25	None
15	None	Dispersion test
16	None	Dispersion test
17	None	Dispersion test
18	None	None
19	None	None
20	None	Dispersion test

Table A2: *(Simulation case 1) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* DHARMa *for the correct model (Model 1, left column) and misspecified model (Model 2, right column) where* $N = 500, \tau = 15, W_i \sim N(1, 2)$ *and* $\bar{D} = 417.75$ *for 20 simulated data sets.*

	Outcome reported by DHARMa	
Run #	Model 1 (correct model, includes W_i)	Model 2 (misspecified model, excludes W_i)
1	None	KS test, dispersion test, outlier test; 0.25
2	None	KS test, dispersion test; 0.25
3	None	KS test, dispersion test; 0.25, 0.75
4	None	KS test, dispersion test, outlier test; 0.25
5	None	KS test, dispersion test, outlier test; 0.25
6	None	KS test, dispersion test; 0.25
7	None	KS test, dispersion test, outlier test; 0.25
8	None	KS test, dispersion test; 0.25
9	None	KS test, dispersion test; 0.25
10	None	KS test, dispersion test; 0.25
11	None	Dispersion test
12	None	Dispersion test; 0.25, 0.75
13	None	KS test, dispersion test; 0.25
14	None	KS test, dispersion test; 0.25
15	None	KS test, dispersion test; 0.25
16	None	KS test, dispersion test; 0.25, 0.75
17	None	Dispersion test, outlier test
18	None	KS test, dispersion test; 0.25, 0.75
19	None	KS test, dispersion test
20	None	Dispersion test, outlier test; 0.25

Table A3: *(Simulation case 1) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* `DHARMa` *for the correct model (Model 1, left column) and misspecified model (Model 2, right column) where* $N = 500, \tau = 7, W_i \sim \mathrm{Bern}(0.5)$ *and* $\bar{D} = 366.85$ *for 20 simulated data sets.*

	Outcome reported by `DHARMa`	
Run #	Model 1 (correct model, includes W_i)	Model 2 (misspecified model, excludes W_i)
1	None	Dispersion test; 0.25
2	None	None
3	None	None
4	None	None
5	None; 0.75	None
6	None	Dispersion test
7	None	KS test, dispersion test; 0.25
8	None	None
9	None	Dispersion test; 0.75
10	None	Dispersion test
11	None	Dispersion test
12	None	Dispersion test
13	None	None
14	None	None
15	None	Dispersion test
16	None	None
17	None	None
18	None	None
19	None	None; 0.25
20	None	Dispersion test

Table A4: *(Simulation case 1) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* `DHARMa` *for the correct model (Model 1, left column) and misspecified model (Model 2, right column) where* $N = 500, \tau = 15, W_i \sim \mathrm{Bern}(0.5)$ *and* $\bar{D} = 424.25$ *for 20 simulated data sets.*

	Outcome reported by `DHARMa`	
Run #	Model 1 (correct model, includes W_i)	Model 2 (misspecified model, excludes W_i)
1	None	Dispersion test; 0.75
2	None	KS test, dispersion test; 0.25, 0.75
3	None	KS test, dispersion test, outlier test; 0.25, 0.75
4	None	Dispersion test; 0.25
5	None	KS test, dispersion test; 0.25
6	None	KS test, dispersion test; 0.25, 0.75
7	None	KS test, dispersion test; 0.25
8	None	KS test, dispersion test; 0.25, 0.75
9	None	KS test, dispersion test; 0.25
10	None	KS test, dispersion test; 0.25
11	None	Dispersion test
12	None	Dispersion test; 0.25, 0.75
13	None	KS test, dispersion test; 0.25
14	None	KS test, dispersion test; 0.25
15	None	KS test, dispersion test; 0.25
16	None	KS test, dispersion test; 0.25, 0.75
17	None	Dispersion test, outlier test
18	None	KS test, dispersion test; 0.25, 0.75
19	None	KS test, dispersion test
20	None	Dispersion test, outlier test; 0.25

Table A5: *(Simulation case 2) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* `DHARMa` *for the correct model (Model 1, left column) and misspecified model (Model 2, right column) where* $N = 500, \tau = 7$ *and* $\bar{D} = 361.10$ *for 20 simulated data sets.*

	Outcome reported by `DHARMa`	
Run #	Model 1 (correct model, includes random effect)	Model 2 (misspecified model, excludes random effect)
1	None	Dispersion test, outlier test
2	None	Dispersion test, outlier test
3	None	Dispersion test; 0.75
4	None; 0.25	Dispersion test, outlier test; 0.25, 0.75
5	Dispersion test	KS test, dispersion test, outlier test; 0.25
6	Dispersion test	Dispersion test, outlier test; 0.25
7	Dispersion test	KS test, dispersion test; 0.25
8	Dispersion test	KS test, dispersion test, outlier test; 0.25
9	None	KS test, dispersion test; 0.25
10	None	KS test, dispersion test; 0.25
11	None	KS test, dispersion test; 0.25, 0.75
12	None	KS test, dispersion test; 0.25, 0.75
13	Dispersion test	Dispersion test; 0.25
14	None	KS test, dispersion test, outlier test; 0.25
15	None	KS test, dispersion test; 0.25
16	None	KS test, dispersion test, outlier test; 0.25
17	None	KS test, dispersion test, outlier test; 0.25, 0.75
18	None	KS test, dispersion test; 0.25, 0.75
19	Dispersion test	Dispersion test
20	Dispersion test	Dispersion test; 0.25, 0.75

Table A6: *(Simulation case 2) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* `DHARMa` *for the correct model (Model 1, left column) and misspecified model (Model 2, right column) where* $N = 500, \tau = 15$ *and* $\bar{D} = 443.80$ *for 20 simulated data sets.*

	Outcome reported by `DHARMa`	
Run #	Model 1 (correct model, includes random effect)	Model 2 (misspecified model, excludes random effect)
1	None	KS test, dispersion test, outlier test; 0.25, 0.50, 0.75
2	None	KS test, dispersion test, outlier test; 0.25, 0.50, 0.75
3	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.50, 0.75
4	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75
5	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75
6	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75
7	None	KS test, dispersion test, outlier test; 0.25, 0.75
8	None	KS test, dispersion test, outlier test; 0.25, 0.75
9	None	KS test, dispersion test, outlier test; 0.25, 0.75
10	None	KS test, dispersion test, outlier test; 0.25, 0.75
11	None	KS test, dispersion test, outlier test; 0.25, 0.75
12	None	KS test, dispersion test, outlier test; 0.25, 0.75
13	Dispersion test	KS test, dispersion test; 0.25, 0.75
14	None; 0.25, 0.50	KS test, dispersion test, outlier test; 0.25, 0.75
15	None	KS test, dispersion test; 0.25
16	None	KS test, dispersion test, outlier test; 0.25
17	None	KS test, dispersion test, outlier test; 0.25, 0.75
18	None	KS test, dispersion test; 0.25, 0.75
19	None	KS test, dispersion test, outlier test; 0.25, 0.50, 0.75
20	None	KS test, dispersion test, outlier test; 0.25

Table A7: *(Simulation case 3) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* `DHARMa` *for the correct model (Model 1, left column) and misspecified model (Model 2, right column) where* $N = 500, \tau = 7$ *and* $\bar{D} = 368.05$ *for 20 simulated data sets.*

	Outcome reported by `DHARMa`	
Run #	Model 1 (correct model, GAM)	Model 2 (misspecified model, GLM)
1	None	None; 0.25, 0.50, 0.75
2	None	None; 0.25, 0.50, 0.75
3	None	None; 0.75
4	None	None; 0.25, 0.50, 0.75
5	None	None; 0.25, 0.50, 0.75
6	None	None; 0.25, 0.50, 0.75
7	None	None; 0.25, 0.50, 0.75
8	None	None; 0.50, 0.75
9	None	None; 0.25, 0.50, 0.75
10	None	Dispersion test; 0.25, 0.50, 0.75
11	None	Dispersion test; 0.25, 0.50, 0.75
12	None	None; 0.25, 0.50, 0.75
13	None	None; 0.25, 0.50, 0.75
14	None	Dispersion test; 0.25, 0.50, 0.75
15	None	None; 0.25, 0.50, 0.75
16	None	None; 0.25, 0.50, 0.75
17	None	None; 0.25, 0.50, 0.75
18	None	None; 0.25, 0.50, 0.75
19	None	None; 0.25, 0.50, 0.75
20	None	None; 0.25, 0.50, 0.75

Table A8: *(Simulation case 3) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* `DHARMa` *for the correct model (Model 1, left column) and misspecified model (Model 2, right column) where* $N = 500, \tau = 15$ *and* $\bar{D} = 434.4$ *for 20 simulated data sets.*

	Outcome reported by `DHARMa`	
Run #	Model 1 (correct model, GAM)	Model 2 (misspecified model, GLM)
1	None	Dispersion test, outlier test; 0.25, 0.50, 0.75
2	None	Dispersion test, outlier test; 0.25, 0.50, 0.75
3	None	Dispersion test, outlier test; 0.25, 0.50, 0.75
4	None	Dispersion test, outlier test; 0.25, 0.75
5	None	Outlier test; 0.25, 0.50, 0.75
6	None	Dispersion test, outlier test; 0.25, 0.50, 0.75
7	None	Dispersion test, outlier test; 0.25, 0.50, 0.75
8	None	None; 0.25, 0.50, 0.75
9	None	None; 0.25, 0.50, 0.75
10	None	Dispersion test; 0.25, 0.50, 0.75
11	None	Dispersion test; 0.25, 0.50, 0.75
12	None	Dispersion test; 0.25, 0.50, 0.75
13	None	Dispersion test; 0.25, 0.50, 0.75
14	None	KS test, dispersion test; 0.25, 0.50, 0.75
15	None	Dispersion test; 0.25, 0.50, 0.75
16	None	Dispersion test; 0.25, 0.50, 0.75
17	None	Dispersion test, outlier test; 0.25, 0.50, 0.75
18	None	Dispersion test, outlier test; 0.25, 0.50, 0.75
19	None	KS test, dispersion test, outlier test; 0.25, 0.50, 0.75
20	None	Dispersion test, outlier test; 0.25, 0.50, 0.75

Table A9: *(Simulation case 4) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* `DHARMa` *for two closed population models, Model 1 (left column) and Model 2 (right column) where* $X \sim N(0, 1)$*, initial* $N = 200$*,* $\tau = 12$*,* $\varphi = \psi = 0.8$ *and* $\bar{D} = 546.50$ *for 10 simulated data sets.*

	Outcome reported by `DHARMa`	
Run #	Model 1 (GLMM)	Model 2 (GLM)
1	Dispersion test; 0.25	KS test, dispersion test, outlier test; 0.25, 0.75
2	KS test, dispersion test; 0.75	KS test, dispersion test, outlier test; 0.25, 0.75
3	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75
4	KS test, dispersion test; 0.75	KS test, dispersion test, outlier test; 0.25, 0.75
5	Dispersion test; 0.75	KS test, dispersion test, outlier test; 0.25, 0.75
6	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75
7	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75
8	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75
9	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75
10	Dispersion test	KS test, dispersion test, outlier test; 0.25, 0.75

Table A10: *(Simulation case 4) Tabulated results comparing the number of goodness-of-fit (GOF) issues reported by* DHARMa *for two closed population models, Model 1 (left column) and Model 2 (right column) where* $X \sim \mathrm{Bern}(0.5)$, *initial* $N = 200$, $\tau = 12$, $\varphi = \psi = 0.8$ *and* $\bar{D} = 577.70$ *for 10 simulated data sets.*

	Outcome reported by DHARMa	
Run #	Model 1 (GLMM)	Model 2 (GLM)
1	Dispersion test; Levene test	KS test, dispersion test, outlier test; Within-group deviation from uniformity
2	None; Levene test	KS test, dispersion test, outlier test; Within-group deviation from uniformity, Levene test
3	Dispersion test	KS test, dispersion test, outlier test; Within-group deviation from uniformity
4	Dispersion test	KS test, dispersion test, outlier test; Within-group deviation from uniformity
5	Dispersion test	KS test, dispersion test, outlier test; Within-group deviation from uniformity
6	Dispersion test; Levene test	KS test, dispersion test, outlier test; Within-group deviation from uniformity
7	Dispersion test; Levene test	KS test, dispersion test, outlier test; Within-group deviation from uniformity, Levene test
8	Dispersion test	KS test, dispersion test, outlier test; Within-group deviation from uniformity
9	Dispersion test	KS test, dispersion test, outlier test; Within-group deviation from uniformity
10	Dispersion test	KS test, dispersion test, outlier test; Within-group deviation from uniformity