



INSECT FEEDING BEHAVIOR STATISTICS

User's Guide

by
ANDERSON RODRIGO DA SILVA
INSTITUTO FEDERAL GOIANO, BRAZIL

JANUARY 2024

Copyright © 2024, Da Silva A.R. All rights reserved.

No part of this material may be reproduced, archived or transmitted, whether electronically, mechanically, by photocopying or otherwise, without written permission from the author.

INFEST® is a free software and comes with ABSOLUTELY NO WARRANTY. For any inquiry or problem, please contact the author/maintainer by e-mail: anderson.silva@ifgoiano.edu.br

The authors would appreciate any suggestions for the improvement of the software or this guide text.

Instituto Federal Goiano - Campus Urutaí
Geraldo S. Nascimento road, km 2.5,
75790-000, Urutaí, Goiás, Brazil
Phone: +55 64 3465-1900

Contents

Introduction	1
Raw data files	2
Data import	3
Processed data tables	5
Exploratory data analysis	6
Modeling with GAMLSS	8
References	14

Introduction

INFEST is a user-friendly web application created to:

- (1) Process multiple EPG recording files containing raw waveform-annotated data.
- (2) Facilitate statistical analysis and modeling.

INFEST was developed in R ([r-project.org](https://www.r-project.org/)) as a backend for processing raw data from EPG systems, tabulation and statistical model fitting. The interface was developed on the frontend (html, css) with the shiny package version 1.3.1 (Chang et al., 2019).

This software was registered at the National Institute of Industrial Property (INPI, Brazil), on May 17, 2022 under the process number BR512022001098-4. It can be used without any licence.

Data tables, figures and analysis reports generated by INFEST can be easily downloaded or transferred to a *MS Office* application.

The current version, 1.2, was released on January 17th, 2024. It can be either installed in a recent (2020 or newer) version of R (<https://www.r-project.org/>) to be downloaded from github.com/arsilva87/infest and ran locally (recommended) or it can run online, from <https://arsilva.shinyapps.io/infest/>, which is 1GB memory-limited, thus being more useful for processing small data sets or quick analyses. New versions of the software are regularly being released on both addresses.

To install INFEST, open up the R console and execute the following command lines:

```
1 if (!require("devtools", quietly = TRUE))
2   install.packages("devtools")
3 devtools::install_github("arsilva87/infest")
```

If needed, let R to update the package dependencies during installation. Once it is finished, execute the following command line to open INFEST:

```
1 infest::infest()
```

Raw data files

INFEST can read raw data by importing two types of files:

- raw data text files (usually **.txt** or **.csv**), containing waveform labels, time and voltage.
- **.ana**, which is structured in three columns: waveform code (integer), time and voltage.

Figure 1 illustrates both types.

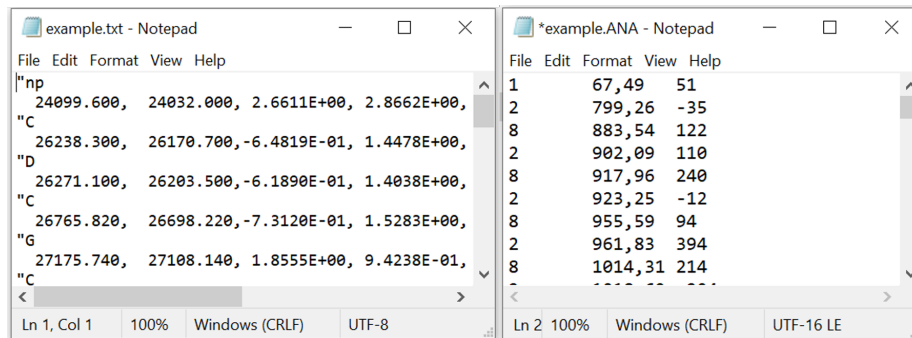


FIGURE 1. Input file types

NOTE: Because INFEST expects characters instead of numbers to identify waveforms, when reading .ana files, the waveform codes are automatically transformed to aphid waveform labels as a standard (Figure 2).

Button	code/F-key	aphid waveform
[np]	1	non penetration
[C]	2	stylet pathway phase, including waveforms A, B, C, and pd
[E1e]	3	E1 extracellular (not always shown)
[E1]	4	E1, normal phloem salivation
[E2]	5	E2, phloem feeding
[F]	6	F, stylet penetration difficulties
[G]	7	G, xylem drinking
[pd]	8	pd, in addition to [C] for separate pd analysis (not recommended)
[II-2]	9	sub-phase II-2 of a pd (if analysed)
[II-3]	10	sub-phase II-2 of a pd (if analysed)
[11]	11	extra button/code for any use (activity event or as a marker)
[12]	12	extra button, as [11]
[T]	99	terminus, total recording time (s) as derived from all hours available in data files. The number 99 can be used in further data processing routines.

FIGURE 2. Aphid waveform codes and labels. Source: extracted from the *Stylet+* manual (<https://www.epgsystems.eu/>)

Data import

From the **Browse** button, one or more files can be imported (Figure 3). There is no need to specify the type of file (.txt, .csv, .ana), as it is automatically recognized and read properly.

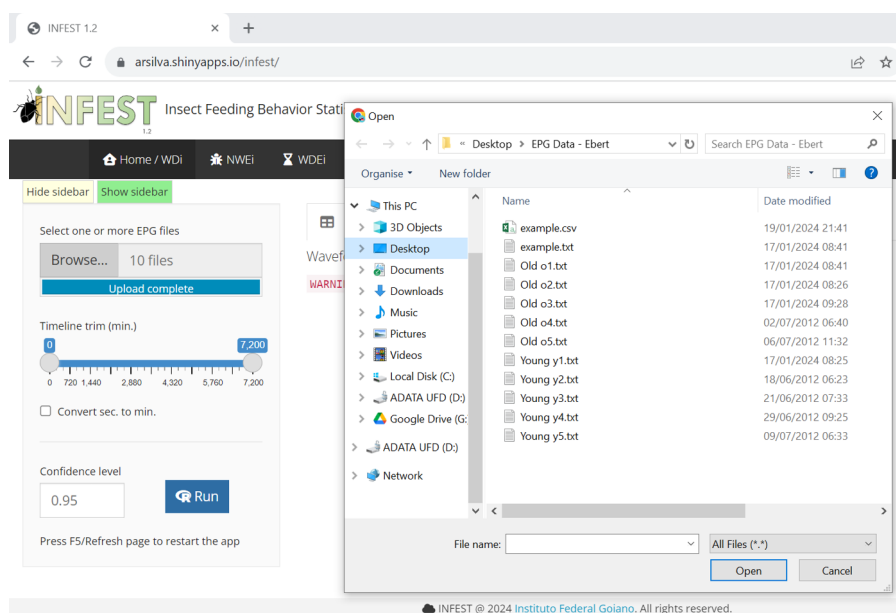


FIGURE 3. Importing multiple raw data files

Concerning raw data file import, INFEST can show two types of messages:

- **Error** (red box): when one or more files are missing one or more waveform labels. In this case, the file(s) is not imported, and the message exhibits which file(s). It requires manual fix by the user.
- **Warning** (yellow box): when there are waveform labels sequentially repeated in one or more files. The message shows which file(s). *NOTE: INFEST does not check for invalid behavior transitions, i.e., illogical sequence of labels.*

Figure 4 illustrates the two types of file import messages.

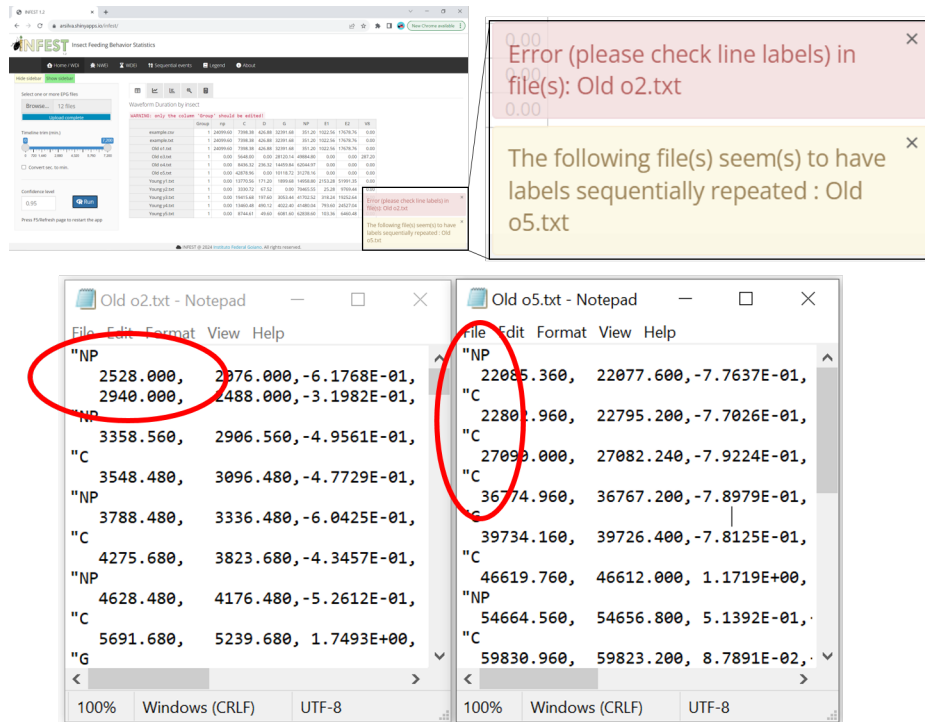



FIGURE 4. Error and warning messages during data import

Processed data tables

After importing files, pressing the Run button generates four processed data tables:






- WDi: Waveform Duration by insect (sec/min). It contains data on the total duration of each waveform event throughout the recordings. The checkbox in the left side menu can be used to convert duration from seconds to minutes (the default).
- NWEi: Number of Waveform Events by insect. It contains data on the total number of waveform events throughout the recordings.
- WDEi: Waveform Duration per Event by insect (sec/min). It consists of the average duration, i.e., the data in WDi divided by their respective data in NWEi.
- Sequential events: a transition matrix containing count data of the type ‘from-to’ of pairwise waveform event sequences. It is recommended to check for valid/invalid behavior transitions.


The output data table WDi in the  ‘Home / WDi’ menu allows the user to edit the column ‘Group’ to specify levels¹ of an experimental factor of interest - for the purpose of statistical analysis. The data tables NWEi and WDEi are automatically updated after any edition in the ‘Group’ column of WDi.

The slide tool ‘Timeline trim’ in the main left side menu can be used to make a time trim (**in minutes only**) at the start or end of the recordings. The maximum value is 7200 minutes, corresponding to 5 days. The mouse or the keyboard arrows can be used to make slide changes. *WARNING: use this tool carefully, as it will cause INFEST to **ignore** the raw data outside the trim limits.*

¹It accepts only integer values to indicate the factor levels.

Exploratory data analysis

The output data tables (WDi, NWEi and WDEi), in their respective panels/menus, are available in the submenu . In each panel, the submenus of icons , ,  and  can be used for data analysis.

In  a principal component biplot is automatically created if the number of observations (input files read) is larger than the number of variables (waveform types). The analysis is based on a correlation matrix, that is, by standardizing variables (waveforms) to have mean zero and unit standard deviation. For example, the insect ‘Young y1.txt’ spent more time in waveforms E1 and E2 than the other insects (Figure 5). ‘Old o2.txt’ is the least active insect (larger values of NP and C). The waveform D contributed less than the others to the variability among insects, as its length over the main axis (Dim1) is relatively smaller. If the column ‘Group’ had been edited and contains levels of a factor, then 95% confidence ellipses² for the mean vector of each level are drawn on the biplot. The plot is interactive and downloadable.

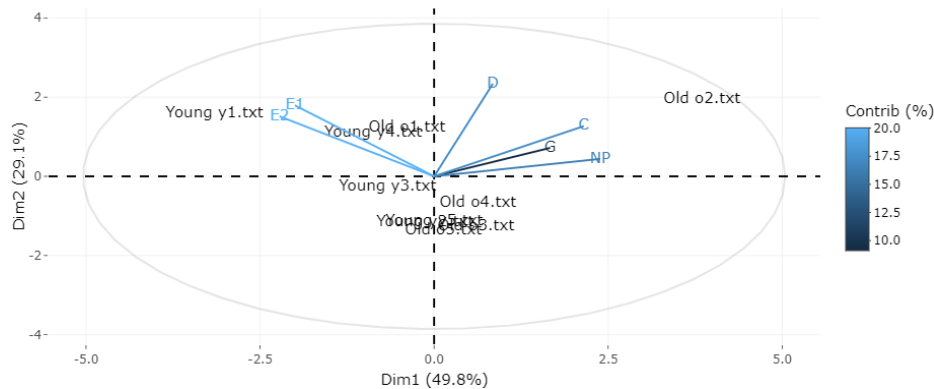





FIGURE 5. PCA biplot for the data table WDi - total duration of waveform events (arrows) by insect (points)

In  a bar-plot for each insect is automatically created. If the column ‘Group’ had been edited and contains levels of a factor, different colors are used to identify the levels. The plot is interactive and downloadable.

In  users can select a response variable (waveform) to create a box-plot from. If the column ‘Group’ had been edited and it contains levels of a factor, multiples box-plots are displayed. It might be useful to detect outliers, analyze the frequency distribution or preview differences among ‘Groups’. The plot is interactive and downloadable.

The submenu  in the ‘Sequential events’ panel can be used to detect invalid transition behaviors in a directional network plot, in which the relative width of

²It can be set from the left side menu.

the arrows are associated with the number of transitions from one waveform to another, considering all insects (input files). Also, the darker (from gray to black) the arrow, the more frequent the transition. The vertex size is also relative and it is associated with the total frequencies of each waveform in table NWEi.

For example (Figure 6), the waveforms C and NP are the most frequent, with plenty of transitions between them, in both directions. It is observed transitions between waveforms E1 and E2. On the other hand, it is not observed a transition from E1 to D.

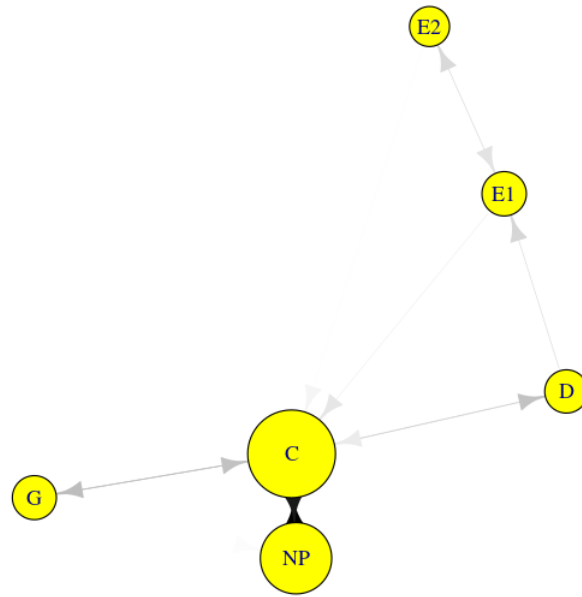



FIGURE 6. Network plot of sequential waveform events (behavior transtions)

Modeling with GAMLSS

The user can fit one-way ANOVA models for a response variable (column) from the data tables WDi, NWEi and WDEi, as a function of the experimental factor specified in the ‘Group’ column in panel .

INFEST’s modeling strategy is based on a very general class of regression models known as GAMLSS - Generalized Additive Models for Location, Shape and Scale (Rigby and Stasinopoulos, 2005), as it allows one to assume for the response variable one out of dozens of probability distributions, including mixtures of distributions and distributions for zero-inflated data.

With GAMLSS, the effect of covariates, such as sex and treatment, can be used to model the mean or median (μ), the scale (σ) and the shape (ν, τ) parameters. For instance, the effect of the level i of a factor ‘Group’ can be used to model μ_i , the expected value of y_i , and σ_i , the true standard deviation of y_i , using the gaussian (normal) distribution, as follows:

$$\begin{aligned}\mu_i &= \beta_i \\ \log(\sigma_i) &= \alpha_i\end{aligned}$$


where β_i and α_i are the effects of level i of ‘Group’ on the mean and standard deviation, respectively; y_i is an observation of the response variable. Now, realize how the parameter σ is related to its predictor (independent variable) through the log function. This is called a *link function*. The link function for μ_i is the *identity*. Another distribution model may have different link functions for each parameter. Table 1 has the default link functions and other main properties of the distribution models available in INFEST, as well as general use indications.

The main assumption in fitting a GAMLSS model is that the observations (y_i) are independent, although it is possible to have correlated random effects in the model.

TABLE 1. Probability distribution models currently available in INFEST to fit GAMLSS

Parameter's default links						
Distribution model	μ	σ	ν	Response interval	Response type	Indication
Gaussian (Normal)	identity	log	-	$-\infty < y < \infty$	Duration and count	Symmetric data
Exponential	log	-	-	$y > 0$	Duration and count	Positively skew data
Gamma	log	log	-	$y > 0$	Duration and count	Positively skew data
Inverse Gamma	log	log	-	$y > 0$	Duration and count	Positively skew data
Inverse Gaussian	log	log	-	$y > 0$	Duration and count	Highly positively skew data
Lognormal	identity	log	-	$y > 0$	Duration and count	Positively skew data
Weibull	log	log	-	$y > 0$	Duration and count	Positively skew data
Pareto type II	log	log	-	$y \geq 0$	Duration and count	Positively skew data
Zero-Adjusted Gamma	log	log	logit	$y \geq 0$	Duration and count	+skew data, $Y = 0$ with probability
Zero-Adj. Inverse Gaussian	log	log	logit	$y \geq 0$	Duration and count	+skew data, $Y = 0$ with probability
Poisson	log	-	-	$y = 0, 1, 2, \dots$	Count	$\text{Var}(Y) = \text{Mean}(Y)$ (equidispersion)
Negative Binomial	log	log	-	$y = 0, 1, 2, \dots$	Count	$\text{Var}(Y) > \text{Mean}(Y)$ (overdispersion)
Zero-Inflated Poisson	log	logit	-	$y = 0, 1, 2, \dots$	Count	EqDisp and $Y = 0$ with probability σ
Zero-Inflated Neg. Binomial	log	log	logit	$y = 0, 1, 2, \dots$	Count	OvDisp and $Y = 0$ with probability ν
Poisson-Inverse Gaussian	log	log	-	$y = 0, 1, 2, \dots$	Count	$\text{Var}(Y) >> \text{Mean}(Y)$


INFEST can assist in selecting the probability distribution model that best fits the data using AIC - Akaike's Information Criterion (lower is better) and RMSE - Root Mean Squared Error (lower is better) as criteria.

To help selecting the best-fitting distribution, a tool in the submenu  was implemented. It consists of an iterative automated fitting of models for y as a function of the intercept only, i.e., with no covariates. So be aware that when fitting a model with covariates such as the factor 'Group', another distribution can fit better the data. But the tool is still a good guide to find a distribution model.

For example, take the total waveform duration by insect in Figure 7. Data³ are given in minutes. The column 'Group' has been edited to specify two levels of an experimental factor.

	Group	NP	C	D	G	E1	E2
Old o1.txt	2	407.51	123.31	7.11	539.86	17.04	294.65
Old o2.txt	2	2272.45	2955.33	9.08	498.30	0.80	0.00
Old o3.txt	2	831.41	98.92	0.00	468.67	0.00	0.00
Old o4.txt	2	1034.08	140.61	3.94	241.00	0.00	0.00
Old o5.txt	2	521.30	714.65	0.00	168.65	0.00	0.00
Young y1.txt	1	355.38	123.44	2.85	31.66	35.89	866.52
Young y2.txt	1	1174.43	55.51	1.13	0.00	0.42	162.82
Young y3.txt	1	695.04	323.59	3.29	50.89	5.30	320.88
Young y4.txt	1	691.33	224.34	8.17	67.04	13.23	408.78
Young y5.txt	1	1047.31	145.74	0.83	101.36	1.72	107.67

FIGURE 7. Waveform Duration by ten insects in two 'Groups'

Let us focus on WDi-C. From Figure 8, the distribution that presented the lowest AIC (139.1) was Inverse Gamma. This suggests that the frequency distribution of C is probably non-symmetric and skewed to the right. This can also be seen in Figure 9, built from the submenu , where, especially for Group 2, the mean (807) is far greater than the median (140.6).

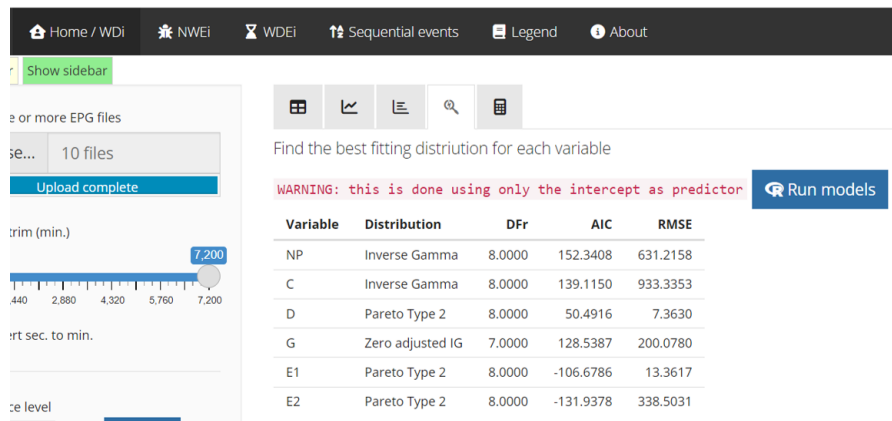


FIGURE 8. Best-fitting distribution for each variable

³This data set was developed by Dr. Timothy Ebert, to whom we are grateful.

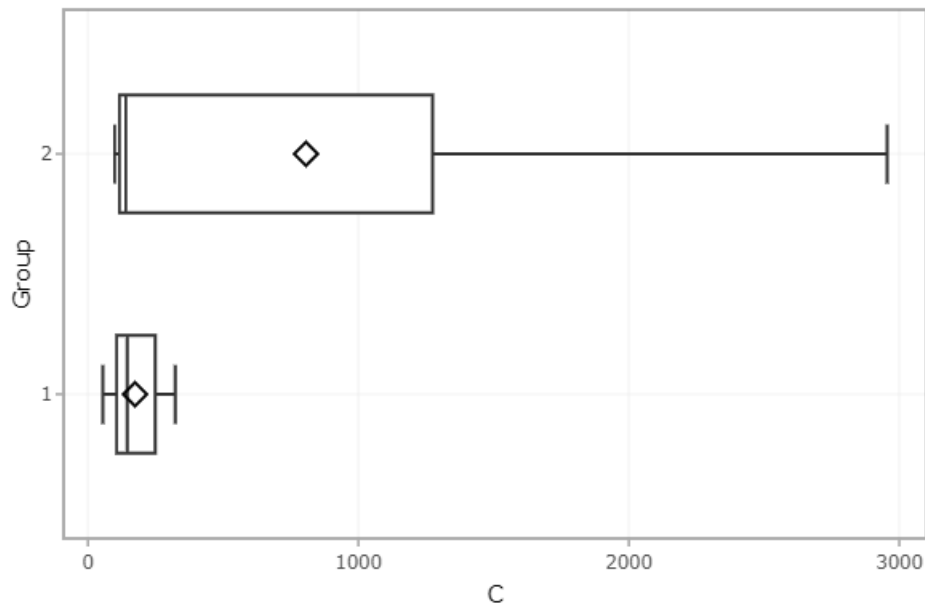


FIGURE 9. Box-plot of the total duration of waveform-C by insect (min)

Notice on Figure 10 how the classical one-way ANOVA model (normal) is having difficulties to detect differences between the Group means (175 *vs.* 807 min), based on the Likelihood Ratio Test's p-value = 0.2174. The model is assuming homoscedasticity, then the estimated standard errors (SE) are the same (349 min), which does not seem to reflect the variability of the observed data (Figure 9).

AIC = 167.5565 RMSE = 779.6819

Likelihood Ratio Test for the 'Group' factor
LRT = 1.52106 df = 1 Pr(Chisq) = 0.2174592

Estimated marginal means

Group	emmean	SE	df	asympt.LCL	asympt.UCL
1	175	349	Inf	-509	858
2	807	349	Inf	123	1490

Confidence level used: 0.95

Pairwise comparisons of means

contrast	estimate	SE	df	z.ratio	p.value
1 - 2	-632	493	Inf	-1.281	0.2001

FIGURE 10. Likelihood ratio test and mean comparisons of WDi-C based on the normal distribution

Now, changing the distribution to Inverse Gaussian and re-fitting the model causes the AIC to fall from 167.5 to 138.9 (Figure 11). Also, the LRT's p-value is

much lower (0.0557). And SE estimates seem to reflect better the variabilities seen in the previous box-plots.

AIC = 138.9377 RMSE = 779.6819

Likelihood Ratio Test for the 'Group' factor
LRT = 3.658111 df = 1 Pr(Chisq) = 0.05579684

Estimated marginal means

Group	response	SE	df	asympt.LCL	asympt.UCL
1	175	59	Inf	90	339
2	807	586	Inf	194	3352

Confidence level used: 0.95
Intervals are back-transformed from the log scale

Pairwise comparisons of means

contrast	ratio	SE	df	null	z.ratio	p.value
1 / 2	0.216	0.173	Inf	1	-1.910	0.0562

Tests are performed on the log scale

FIGURE 11. Likelihood ratio test and mean comparisons of WDi-C based on the Inverse Normal distribution

Pairwise mean comparisons are done using a normal approximation, based on the z-score. Tukey's method can be used (mark the submenu checkbox) to correct type-I error rate due to multiple comparisons.

A graphical tool to make diagnostics on the fitted model is a worm-plot, which is similar to a Q-Q plot in both construction and interpretation, but using detrended residuals instead. Then, deviations from the residual mean (horizontal straight line on zero) outside the (95%) confidence limits (dashed lines) indicate fitting problems at some region of the predictor values. Figure 12 shows that the fitted model based on Inverse Normal does not present such problems.

ADVICE:

Since there are many distribution models available in INFEST, it is strongly recommended to **compare the results from several models, always**.

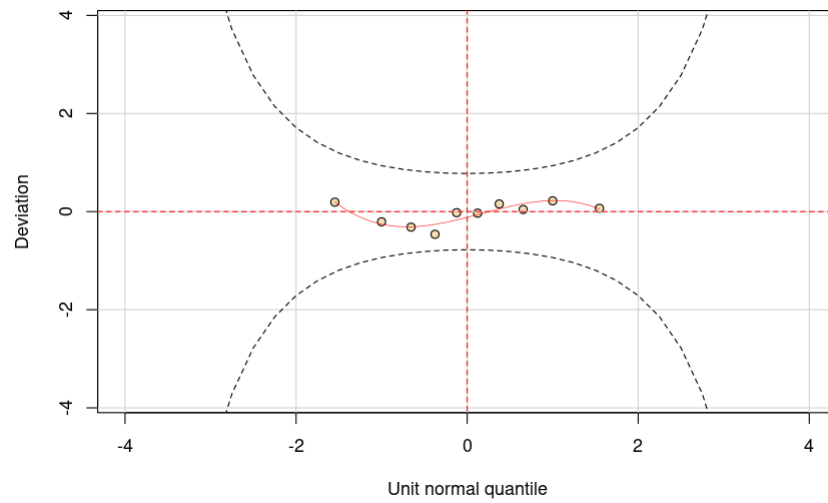


FIGURE 12. Worm-plot for model fitting diagnostics

References

- [1] Chang, W., Cheng, J., Allaire, J.J., Xie, Y., McPherson, J. (2019). shiny: Web Application Framework for R. R package version 1.3.1. Available from: <https://CRAN.R-project.org/package=shiny>
- [2] Lenth, R. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.3. <https://CRAN.R-project.org/package=emmeans>
- [3] R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- [4] Rigby, R. A., and D. M. Stasinopoulos (2005). Generalized Additive Models for Location, Scale and Shape. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54 (3): 507-54.
- [5] Silva, A. R. da, Almeida, A. C. S., Gonçalves de Jesus, F., Barrigossi, J. A. F. (2024). INFEST: Insect Feeding Behavior Statistics. INPI - INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL [BR512022001098-4]. Available at: arsilva.shinyapps.io/infest, <https://github.com/arsilva87/infest/>