Supplement 8 - R-squared measures for linear models fitted to zero-inflated prey capture data

2025-03-01

Introduction

Studies aiming to assess how well wiggles (or other behavioural proxies derived from bio-logging data) predict prey captures often use linear regression to relate observed prey capture events to wiggle (proxy) counts. This approach evaluates the strength and direction of the relationship, offering a simple and intuitive measure of how closely the proxy tracks true values (e.g., Takahashi et al. 2004, Bost et al. 2007).

However, prey capture data are count-based and typically exhibit zero inflation, with more zeros than expected under standard count distributions. From a statistical perspective, linear regression is not ideal, as it assumes normally distributed errors. Despite this limitation, linear models remain a commonly used tool for preliminary analyses due to their simplicity and interpretability.

R-squared values derived from linear models are often used to quantify how much of the variation in observed prey captures is explained by the proxy variable (e.g., wiggle counts). The simple example below demonstrates how "zero-zero" observations and predictions can inflate R-squared values derived from linear models, which fail to account for the distributional properties of count data.

Simulate data (No. of prey captures and wiggles per dive)

```
library(tidyverse)

set.seed(1234)  # for reproducibility

# Simulate video prey captures (the total number of prey captures per dive):
# 30 random integers between 0 and 30
video_prey_captures <- sample(0:30, 30, replace = TRUE)

# Create wiggle counts with some linear correlation + noise
wiggle_counts <- round(video_prey_captures * 1.5 + rnorm(10, mean = 5, sd = 25))

# Ensure wiggle_counts stay >= 0
wiggle_counts[wiggle_counts < 0] <- 0

# Create data tibble
data <- tibble(video_prey_captures = video_prey_captures,
               wiggle_counts = wiggle_counts)

# Add 10 "zero-zero" observations (non-foraging dives, correctly predicted)
data <- data %>%
    add_row(video_prey_captures = rep(0, 10),
            wiggle_counts = rep(0, 10))
```

```r
# The above data are mostly dives with prey captures. But there can be many dives
# with zero prey captures and zero predictions. There are, for example,
# travelling dives or other exploratory dives (V-shaped dives).
# Now make the above data zero-inflated by adding another 20 zeros to both columns.
# This represents 20 dives with no prey captures, and no wiggles.

data_zi <- data %>%
    add_row(video_prey_captures = rep(0, 20),
            wiggle_counts = rep(0, 20))
```

Fit linear models and assess R-squared

```r
# Fit a linear model to data
linear_model = lm(video_prey_captures ~ wiggle_counts, data = data)

# Get R-squared
r_squared <- summary(linear_model)$r.squared
r_squared
```

```
## [1] 0.2504552
```

```r
# Fit linear model to zero-inflated data (typical scenario)
linear_model_zi = lm(video_prey_captures ~ wiggle_counts, data = data_zi)

# Get R-squared
r_squared_zi <- summary(linear_model_zi)$r.squared
r_squared_zi
```

```
## [1] 0.3756741
```

This simple example illustrates that the presence of numerous non-foraging dives — correctly predicted as such — can inflate the R-squared value obtained from a linear regression. However, correctly identifying numerous non-foraging dives - with a resultant 'high' R-squared value - is arguably of limited value if predictions for foraging dives show high variability around the regression line.

Thus, zero inflation can add "easy wins" - i.e., many observed zeroes that are easy to predict. Predicting these well adds strongly to the model's apparent ability to explain the variation in the variable of interest, but this doesn't mean the model predicts the variation in counts well. One should thus be mindful not to attach too much importance to R-squared values obtained from similar data.

References

Bost, C.A., Handrich, Y., Butler, P.J., Fahlman, A., Halsey, L.G., Woakes, A.J., & Ropert-Coudert, Y. 2007. Changes in dive profiles as an indicator of feeding success in king and Adélie penguins. Deep Sea Research Part II: Topical Studies in Oceanography 54: 248-255.

Takahashi, A., Dunn, M.J., Trathan, P.N., Croxall, J.P., Wilson, R.P., Sato, K., Naito, Y. 2004. Krill-feeding behaviour by a chinstrap penguin Pygoscelis antarctica compared to fish-eating in magellanic penguins Spheniscus magellanicus: a pilot study. Marine Ornithology 32: 47-54.