Kato-Katz Infection Detection Probability Model

# What the Project Does

This project uses open-source data obtained from the peer-reviewed article "Prevalence of Schistosoma mansoni and Soil-Transmitted Helminths in a Rural Population of the Philippines: A Cross-Sectional Study", specifically EPG data from individuals infected with Schistosoma mansoni (S.m.). The model estimates the true prevalence of infection in populations diagnosed with a single Kato-Katz (1KK) slide and predicts how the prevalence would change if multiple Kato-Katz slides (4KK) were used instead. This approach addresses the challenge of resource limitations by modeling the relationship between infection intensity (EPG) and detection probability, providing insights into how many infections may have been missed and helping inform more accurate public health decisions and prevalence estimates.

 **Linear Model (model\_logit)**:  
The first model uses a **linear regression (lm)** approach to model the relationship between the logit-transformed detection probability and the observed egg count (EPG) for a single Kato-Katz (KK) test. The equation is:

Ratio\_logit=β0+β1⋅KK\_1\_EPG

This model assumes a linear relationship between the log-transformed detection probability and the raw EPG values. It provides a straightforward way to estimate the detection probability but may not fully capture any non-linearities in the data.

 **Linear Model with Square Root Transformation (model\_logit\_lm)**:  
The second model also uses **linear regression**, but it transforms the **KK\_1\_EPG** by applying the square root function. This transformation is often used in ecological and epidemiological data to stabilize variance and linearize relationships between variables that may have a non-linear relationship.

The equation for this model is:

Ratio\_logit=β0+β1⋅sqrt(KK\_1\_EPG)

By using the square root of the EPG value as the predictor, this model accounts for the fact that the detection probability may increase more slowly as the egg count rises. It may better capture the behavior of the infection intensity in relation to detection.

 **Generalized Linear Model (model\_glm)**:  
The third model uses a **binomial generalized linear model (GLM)** with a logit link. It is designed to model the probability of a positive result in the Kato-Katz test as a function of the square root of the **KK\_1\_EPG**. This model is more appropriate when the outcome variable is binary (positive vs. negative result), and it allows for a direct estimate of the probability of detection for each individual.

The equation for this model is:

logit(P(positive))=β0+β1⋅KK\_1\_EPG

By using a binomial family and the logit link, this model works directly with the proportion of positive test results, rather than transforming the outcome through the logit.

 **Generalized Additive Model (model\_gam)**:  
The final model is a **Generalized Additive Model (GAM)**. This model is more flexible and can capture non-linear relationships between the infection intensity and detection probability. Unlike the linear models, the GAM allows the predictor **KK\_1\_EPG** to enter the model through a smooth function, denoted by s(KK\_1\_EPG)s(KK\_1\_EPG). This is useful when there is an expected non-linear relationship between the egg count and detection probability, which might be missed by the linear models.

The equation for this model is:

Ratio\_logit=β0+s(KK\_1\_EPG)

The smooth function s(KK\_1\_EPG)s(KK\_1\_EPG) is estimated by the model, which allows it to fit more complex patterns in the data.

# Why the Project is Useful

# This project helps estimate the ‘true’ prevalence of infection in populations where only a single Kato-Katz (1KK) slide is used for diagnosis, rather than the more accurate multiple slide (4KK) method. In resource-limited settings, where conducting multiple KK slides may not be feasible, this model predicts how many infections might be present in the population with only using 1 KK, helping to improve prevalence estimates. By understanding the potential underestimation of cases detected by 1KK, this model can guide more informed public health decisions and resource allocation for interventions.

# How to Use the Model

The model uses the intensity of infection(EPG) observed from 1KK, using pre determined probabilities of a single KK testing positive at a given EPG for S.m. This EPG is used to predict a logit transformed probability. Users can run these models on their own data by simply using a .csv of EPG values to then predict the population level prevalence.

# Where to Get Help

For support or to report issues, please open an issue on the repository's GitHub page: [GitHub Issues](https://github.com/WilliamJonesWarner/ S.Mansoni\_Population\_prediction/issues).

# Maintainers and Contributors

- Maintained by: William Jones-Warner - Contributions are welcome! Please fork the repository and submit a pull request with your improvements or bug fixes.