SHW PGT METHODS MASTERCLASS

Beyond the Log Transformation

Modelling Zero-Inflated Continuous (Semi-Continuous) Data Using Tweedie Generalised Linear Mixed Models (GLMMs)

A practical using synthetic Scottish Health Survey (SHeS) data

Shaofen Xu | PhD Candidate in Social and Public Health Sciences

Health Economics and Health Technology Assessment Unit, School of Health and Wellbeing
University of Glasgow

The Data Problem

Spike at Zero

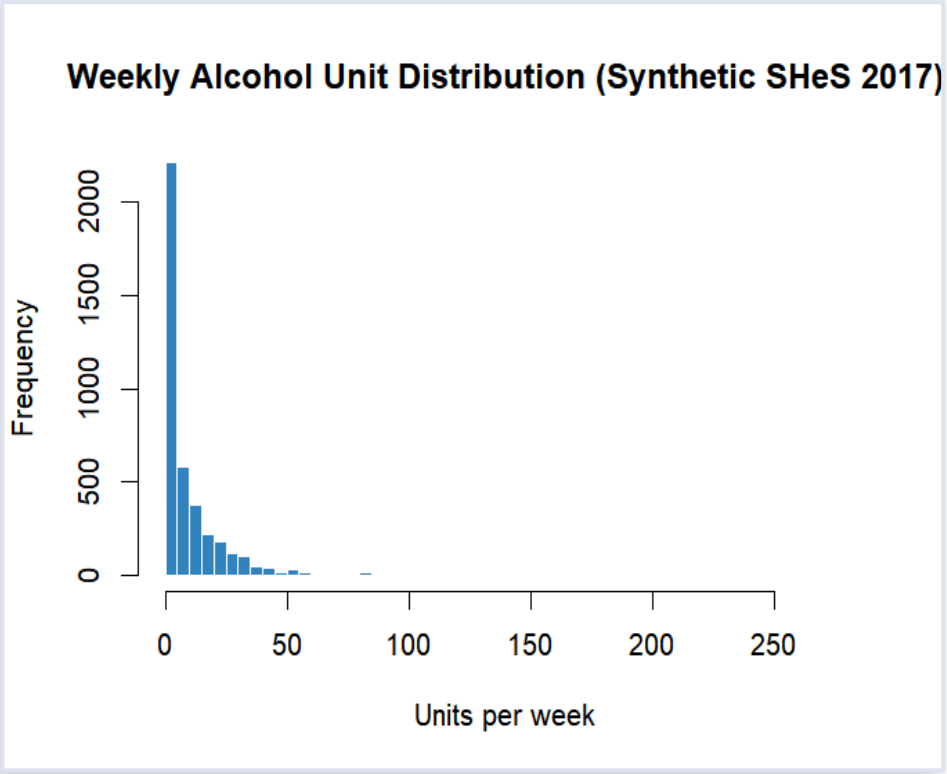
Many respondents report zero consumption — absolute, not missing.

Right-Skewed Tail

Those who do drink span a wide range;
A few consume very large quantities.

Continuous Scale

Values like 0.7, 20.8, 50.1 alcohol units — not integers.



Why Traditional Approaches Fail

✗ **Standard Linear Regression (Ordinary Least Squares)**

Assumes normally distributed residuals.
Right-skewed zero-inflated data violates this immediately.
Predicted values can be negative.

✗ **Log-transformation of raw data — log(y + 1)**

Adding 1 is an arbitrary choice that distorts the outcome.
Produces funnel-shaped residuals (heteroscedasticity).
Back-transformation via “exp()” is biased: $E[\exp(\log Y)] \neq E[Y]$.

✗ **Zero-Inflated Count Models (e.g., Zero-Inflated Negative Binomial, Zero-inflated Poisson)**

Designed for integers — hospital visits, event counts.
UK alcohol units are continuous decimals.
Applying a count model to non-integer data is mathematically invalid.

Enter the Tweedie Distribution

A compound Poisson-Gamma process — two behaviours, one model

Poisson Component

Whether to drink at all this week?

Outputs integers/counts like zeros

+

Gamma Component

Given drinking occurred, how much?

Outputs continuous right-skewed positive values

=

Tweedie Distribution

Handles zeros
AND
Continuous right-skewed positive data

Single model · No data splitting · One set of coefficients · One variance structure

The Power Variance (Tweedie) Family

Tweedie is actually a big family: Normal (Gaussian), Poisson, Gamma and inversed Gaussian distributions.
All share the mean-variance relation: $Var = \phi \cdot \mu^p$. **p (psi or ψ) stands for variance power.**
This family and other distributions (e.g., Negative Binomial) constitute Exponential Dispersion Family (= Generalised Linear Models).
Generalised Linear Mixed Models (GLMMs) are multilevel/multivariable Generalised Linear Models (GLMs).

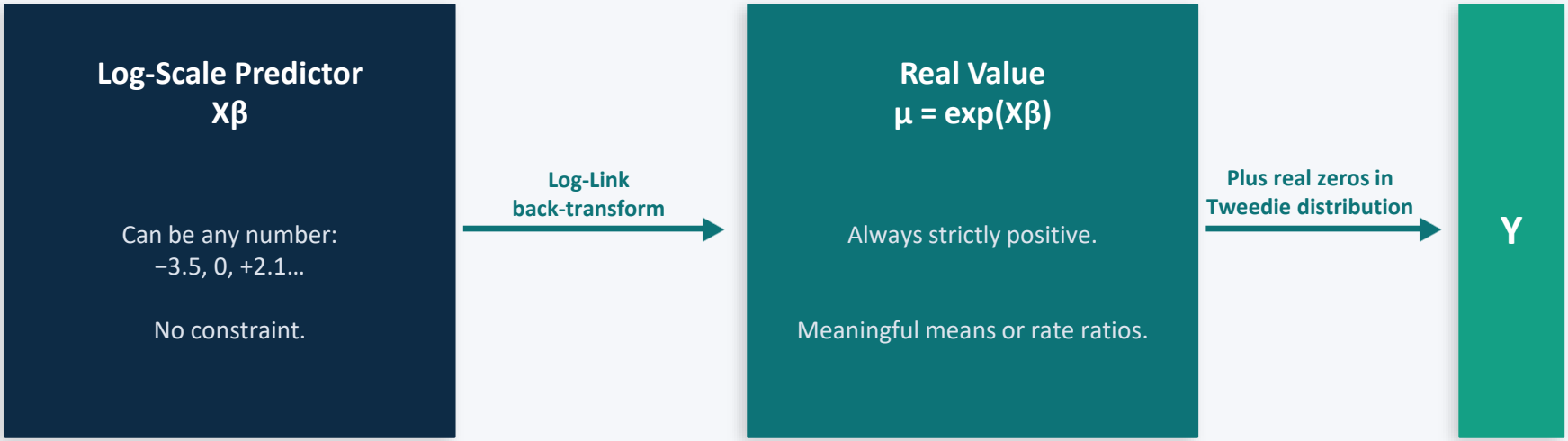
Undefined (0 < p < 1)	OUR ZONE 1 < p < 2	No one use in real-world (2 < p < 3)

p = 0	p = 1	p = 2	p = 3
Normal	Poisson	Gamma	Inv. Gaussian

Normal (p = 0)	Poisson (p = 1)	Tweedie distribution (1 < p < 2)	Gamma (p = 2)
Constant variance	Var = Mean	Flexible mean-variance relation	Var ∝ Mean²
Standard linear regression	Count/integer data	Right-skewed semi-continuous	Right-skewed positive continuous

99% Scenarios: Tweedie + Log-Link

Do not log the raw data. Log the link. Use the build-in function of packages.



Log-link guarantees the predicted value is only positive.
The **Tweedie distribution** then add actual observed zeros.
They operate at different levels but come together in almost any real-world analyses.

Why Not a Hurdle (Two-Part) Model under Frequentist?

Hurdle / Two-Part Model

Part 1: Logistic regression
→ Odds ratios and predicted probabilities of event occurrence

Part 2: Gamma/Lognormal regression
→ Rate ratios and means when events occurred

Two separate parts
Two sets of coefficients

In GLMMs (e.g., intersectional MAIHDA):
Doubles the coefficient sets and random effect structures

Likely covariance and no model convergence
Result interpretation is fragmented into two scenarios

Tweedie MAIHDA GLMM

ONE modelling part
ONE go for intersectional strata (can be many!)

ONE set of coefficients
ONE set of random effects

No covariance between random effect variance structures

Stable model convergence for each stratum

Clean, interpretable results without conditional scenarios

Hands-On: Your Coding Environment

Dataset

"SHeS_Masterclass_Synthetic.csv"

Generated using "synthpop":
preserves the joint distribution, zero proportion, and right-skewed tail of the real 2017 SHeS cohort.

Real SHeS cannot be shared under relevant data governance.

Key R Packages

"glmmTMB"
— Tweedie MAIHDA GLMM fitting

"easystats"
— model_parameters()
— icc() for variance/inequality structures
— compare_parameters()

Install according to your:
"Tweedie_Masterclass_Script.R"

Download Data and Script

https://github.com/shaofenxu/Tweedie-GLMM-Practical.git

LIVE CODING PRACTICAL

Switch to RStudio

01 **Explore the Data**

hist(shes_data$drating) — confirm zero spike and right tail

02 **Null MAIHDA Model**

m_null: extract VPC via icc() — how much variance sits between strata?

03 **Full MAIHDA Model**

m_main: add Sex, age_group, SIMD — compare_parameters(), VPC, PCV

04 **Check Variance Power p**

Confirm 1 < p < 2 — this parameter tells if the distribution is correctly chosen

05 **Interpret Predicted Values**

model_parameters(m_main, exponentiate = TRUE) — Rate Ratios and Means

THE HONEST TRADE-OFF

Tweedie is elegant — but not perfect.

Key Limitation

Tweedie assumes the probability of the outcome occurring and its intensity are driven in the same direction by every covariate. If occurrence increased but intensity decreased (or vice versa), Tweedie will average over these opposing effects and miss both mechanisms. In that situation, a Hurdle model provides more nuanced insights.

Check your research question

Focus on population averages? Use Tweedie: no scenario.
Focus on behavioural mechanisms? Use Hurdle: two scenarios.

Questions? [x.xu.2@research.glasgow.ac.uk]

References

Traditional log-transformation — log(y + 1):
SHMUELI, G., JANK, W. & HYDE, V. 2008. Transformations for semi-continuous data. *Computational Statistics & Data Analysis, 52**, 4000-4020.***

Zero-inflated count models:
SALEHI, M. & ROUDBARI, M. 2015. Zero inflated Poisson and negative binomial regression models: application in education. *Med J Islam Repub Iran, 29**, 297.***

Exponential dispersion family — GLMs including GLMMs:
JØRGENSEN, B. 1992. Exponential Dispersion Models and Extensions: A Review. *International Statistical Review / Revue Internationale de Statistique, 60**, 5-20.***
BINGHAM, N. H. & FRY, J. M. 2010. Generalised Linear Models. *In: BINGHAM, N. H. & FRY, J. M. (eds.) Regression: Linear Models in Statistics. London: Springer London.*
GOLDSTEIN, H. 1991. Multilevel Modelling of Survey Data. *Journal of the Royal Statistical Society. Series D (The Statistician), 40**, 235-244.***

Applications of Tweedie distribution + log-link:
SHONO, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research, 93**, 154-162.***
QUIJANO XACUR, O. A. & GARRIDO, J. 2015. Generalised linear models for aggregate claims: to Tweedie or not? *European Actuarial Journal, 5**, 181-202.***

Hurdle (two-part) models:
FENG, C. X. 2021. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications, 8**, 8.***
LIU, L., SHIH, Y.-C. T., STRAWDERMAN, R. L., ZHANG, D., JOHNSON, B. A. & CHAI, H. 2019. Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review. *Statistical Science, 34**, 253-279.***

R practical and data:
NOWOK, B., RAAB, G. M. & DIBBEN, C. 2016. synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software, 74**, 1 - 26.***
Lüdecke, D., Ben-Shachar, M.S., Patil, I., Wiernik, B.M., Bacher, E., Thériault, R. and Makowski, D. (2022) 'easystats: Framework for Easy Statistical Modeling, Visualization, and Reporting', *CRAN*. Available at: https://doi.org/10.32614/CRAN.package.easystats.
Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Mächler, M. and Bolker, B.M. (2017) 'glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling', *The R Journal*, 9(2), pp. 378–400. Available at: https://doi.org/10.32614/RJ-2017-066.
Hartig, F. (2024) *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models* (Version 0.4.7) [Computer software]. Available at: https://CRAN.R-project.org/package=DHARMa.
ScotCen Social Research, 2021, *Scottish Health Survey, 2017*, [data collection], UK Data Service, Accessed 1 May 2026. SN: 8398, DOI: http://doi.org/10.5255/UKDA-SN-8398-1