ARTIFICIAL EXAMPLE 1:

INTERPRETATION OF PARAMETERS IN THE ZERO-PART OF THE ZERO-INFLATED POISSON MODEL

Suppose we want to assess the effect of binary covariate x_{1i}, (e.g., education level being high or low) on a count variable Y_i (e.g., counting the number of UPB-perpetrations), and assume that Y_i follows a zero-inflated Poisson distribution.

We generate data following models (1), (2), (5) and (6) of the paper (with $\boldsymbol{x}_i^t \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i}$ and $\boldsymbol{x}_i^t \boldsymbol{\gamma} = \gamma_0 + \gamma_1 x_{1i}$) and consider the following hypothetical values for the parameters: β_0 = log 2, β_1 = log 2, γ_0 = log (log 2) and γ_1 = log (log 6 / log 2). For these specific choices it can easily be shown that $\Pr(Y_i = 0 \mid X_{1i} = 0)$ equals $\Pr(Y_i = 0 \mid X_{1i} = 1) = 83\%$.In other words the proportion of subjects with a zero count does not depend on the level of x_1. As a consequence, the parameter β_1^* in hurdle model (7), which captures the effect of x_1 on the zero counts, equals 0. Figure 1 shows for a simulated sample of size 1000 under such scenario the observed distribution of Y for the separate levels of x_1. The proportion of zero counts is indeed about equal for the two levels of x_1.

Looking at the fitted ZIP-model and the estimated effect of β_1 (table 1) we find that the estimated odds of 'excess zeros' is about exp(1.06) = 2.85 (95% CI: 1.50 to 5.50) times higher in the x_1 = 1-group than in the x_1 = 0-group (p = 0.001), which may lead to the erroneous interpretation that the odds of observing zero counts is significantly larger in the x_1 = 1-group than in the x_1 = 0-group. The latter can only be derived directly from the hurdle model that cleanly separates the zero counts and non–zero counts. From the left lower panel of table 1 we indeed observe no effect of x_1 in the logistic part of the hurdle model as the proportion of observed zeros is approximately equal between both levels of x_1. (Note that in the table the signs of the parameters in the zero-component of the hurdle model are reversed compared to the R-output, as we have chosen to model the probability of a zero count instead of the non-zero counts.)

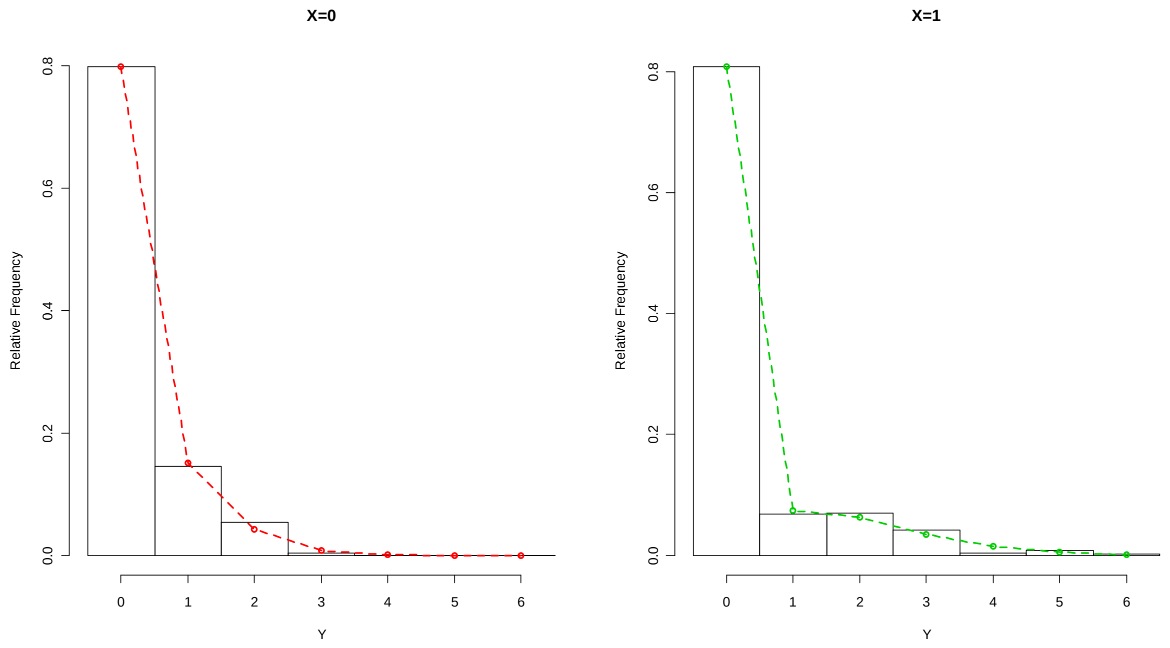

Figure 1: Empirical and Estimated Count Distribution by X-level

	Logistic portion			Counts portion		
Variable	β	SE β	Z	γ	SE γ	Z
			ZIP-model			
Intercept	0.11	0.31	0.37	-0.58	0.18	-3.26**
x_1	1.06	0.33	3.20**	1.10	0.20	5.51***
			PLH-model			
Intercept	1.37	0.11	11.95***	-0.58	0.18	-3.26**
x_1	0.06	0.16	0.38	1.10	0.20	5.51***

Table 1: Hypothetical Example 1: Estimated Parameters under Zero-Inflated Poisson Model and Poisson Logit Hurdle Model. (*** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$)