How Many People Visit YouTube? Imputing Missing Events in Panels With Excess Zeros

Georg M. Goerg, Yuxue Jin, Nicolas Remy, Jim Koehler[1]

[1] Google, Inc.; United States

E-mail for correspondence: `gmg@google.com`

Abstract: Media-metering panels track TV and online usage of people to analyze viewing behavior. However, panel data is often incomplete due to non-registered devices, non-compliant panelists, or work usage. We thus propose a probabilistic model to impute missing events in data with excess zeros using a negative-binomial hurdle model for the unobserved events and beta-binomial sub-sampling to account for missingness. We then use the presented models to estimate the number of people in Germany who visit YouTube.

Keywords: imputation; missing data; zero inflation; panel data.

1 Introduction

Media panels (GfK Consumer Panels, 2013) are used by advertisers to estimate *reach* and *frequency* of a campaign: reach is the fraction of the population that has seen an ad, frequency tells us how often they have seen it (on average). It is important to get good estimates from panel data, as they largely determine the cost of an ad spot on TV or a website.

Naïvely, one would use a sample fraction of the number of non-zero events (website visits, TV spots watched, etc.) per unit time to estimate reach; similarly, for frequency. This, however, suffers from underestimation as panels often only record a fraction of all events due to e.g., non-compliance or work usage. Correcting this bias and imputing missing events has been studied previously (Fader and Hardie, 2000; Yang et al., 2010).

In this work we i) extend the beta-binomial negative-binomial (BBNB) model (Hofler and Scrogin, 2008) with a hurdle component to improve modeling excess zeros in panel data (§2); ii) present the maximum likelihood estimator (MLE) and also add prior information on missingness (§3); and iii) use the methodology to estimate – from online media panels and internal YouTube log files – how many people in Germany visit YouTube (§4).

The proposed methodology can be applied to a great variety of situations where events have been counted – but some are known to be missing.

2 Hierarchical Event Imputation

Let $N_i \in \{0, 1, 2, \ldots\}$ count the true (but unobserved) number of visits by panelist i. The population consists of people who do not visit YouTube at all (with probability $q_0 \in [0, 1]$), and those who visit at least once. If she visits (overcoming the "hurdle" with probability $1 - q_0$), we assume that N_i is distributed according to a shifted Poisson distribution (starting at $n = 1$) with rate λ_i. For model heterogeneity among the population we use a Gamma $\left(r, \frac{q_1}{1-q_1}\right)$ prior for λ_i, with $r > 0$ and $q_1 \in (0, 1)$.
Overall, this yields a shifted negative binomial hurdle (NBH) distribution

$$\mathbb{P}\left(N = n; q_0, q_1, r\right) = \begin{cases} q_0, & \text{if } n = 0, \\ (1 - q_0) \cdot \frac{\Gamma(n+r-1)}{\Gamma(r)\Gamma(n)} \cdot (1 - q_1)^r q_1^{n-1}, & \text{if } n \geq 1. \end{cases} \quad (1)$$

We choose a hurdle, rather than a mixture, model for the excess zeros (Hu et al., 2011), since $1 - q_0$ can be directly interpreted as the true – but unobserved – 1+ reach: if an advertiser shows an ad on YouTube they can expect that a fraction of $1 - q_0$ of the population sees it at least once.
Let p_i be the probability a visit of user i is recorded in the panel. Assuming independence across visits the total number of recorded panel events, $K_i \in \{0, 1, 2, \ldots\}$, thus follows a binomial distribution, $K_i \sim \text{Bin}(N_i, p_i)$. To account for heterogeneity across the population we assume $p_i \sim \text{Beta}(\mu, \phi)$, with mean μ and precision ϕ (Ferrari and Cribari-Neto, 2004). Here μ represents the expected non-missing rate and ϕ the (inverse) variation across the population. Integrating out p_i gives a Beta-Binomial (BB) distribution,

$$K_i \mid N_i \sim BB(N_i; \mu, \phi). \quad (2)$$

Combining (1) and (2) yields a hierarchical beta-binomial negative-binomial hurdle (BBNBH) imputation model with parameter vector $\theta = (\mu, \phi, q_0, r, q_1)$:

$$N_i \sim NBH(N; q_0, r, q_1) \text{ and } K_i \mid N_i \sim BB(K \mid N_i; \mu, \phi). \quad (3)$$

2.1 Joint Distribution

The pdf of (2) can be written as

$$g(k \mid n; \mu, \phi) = \binom{n}{k} \frac{\Gamma(k + \phi\mu)\Gamma(n - k + (1 - \mu)\phi)}{\Gamma(n + \phi)} \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1 - \mu))} .$$

For $k = 0$ this reduces to

$$\mathbb{P}\left(K = 0 \mid N, \mu, \phi\right) = \frac{\Gamma(n + (1 - \mu)\phi)}{\Gamma(n + \phi)} \times \frac{\Gamma(\phi)}{\Gamma(\phi(1 - \mu))} . \quad (4)$$

Due to the zero hurdle it is useful to treat $N = 0$ and $N > 0$ separately:

$$\mathbb{P}\left(N, K\right) = \mathbb{P}\left(K \mid N\right) \cdot \mathbb{P}\left(N\right) = BB(k \mid n; \mu, \phi) \cdot NBH(n; q_0, q_1, r) \qquad (5)$$

For $n = 0$, (5) is non-zero only for $k = 0$, $\mathbb{P}\left(N = 0, K = 0\right) = q_0$, since $\mathbb{P}\left(K > N\right) = 0$. For $n > 0$,

$$\begin{aligned}\mathbb{P}\left(N = n, K = k\right) = &(1 - q_0)\frac{1}{B(\phi\mu, \phi(1-\mu))}\frac{(1-q_1)^r}{\Gamma(r)} \times \frac{\Gamma(k + \phi\mu)}{\Gamma(k+1)} \\ &\times \frac{\Gamma(n - k + \phi(1-\mu))}{\Gamma(n-k+1)}\frac{\Gamma(n + r - 1)}{\Gamma(n + \phi)}q_1^{n-1} \times \frac{\Gamma(n+1)}{\Gamma(n)} .\end{aligned} \qquad (6)$$

2.2 Conditional Predictive Distribution For Imputation

The panel records k_i events for panelist i, but we want to know how many events truly occurred. That is, we are interested in (dropping subscript i)

$$\mathbb{P}\left(N = n \mid K = k\right) = \frac{\mathbb{P}\left(K = k \mid N = n\right)\mathbb{P}\left(N = n\right)}{\mathbb{P}\left(K = k\right)}, \qquad (7)$$

To obtain analytical expressions we consider $k = 0$ and $k > 0$ separately:

$k = 0$: Either none truly happened ($n = 0$) or a panelist visited at least once ($n > 0$), but none were recorded.

$n = 0$:

$$\mathbb{P}\left(N = 0 \mid K = 0\right) = \frac{q_0}{\mathbb{P}\left(K = 0\right)}. \qquad (8)$$

$n > 0$:

$$\begin{aligned}\mathbb{P}\left(N = n \mid K = 0\right) = &\frac{1}{\mathbb{P}\left(K = 0\right)} \times \frac{\Gamma(n + \phi(1-\mu))}{\Gamma(n + \phi)}\frac{\Gamma(\phi)}{\Gamma(\phi(1-\mu))} \\ &\times (1 - q_0)\frac{\Gamma(n + r - 1)}{\Gamma(n)}\frac{(1-q_1)^r}{\Gamma(r)}q_1^{n-1},\end{aligned}$$

where the second term comes from (4).

$k > 0$: The zero "hurdle" for N has been surpassed for sure.

$n < k$: By construction of Binomial subsampling

$$\mathbb{P}\left(N = n \mid K = k\right) = 0 \text{ for all } n < k. \qquad (9)$$

$n \geq k$: Here

$$\mathbb{P}\left(N = n \mid K = k\right) = n \cdot q_1^{n-1}\frac{\Gamma(n - k + (1-\mu)\phi)}{\Gamma(n-k+1)\Gamma(n+\phi)}\Gamma(n + r - 1) \times$$
$$\left(\sum_{m=0}^{\infty}(m + k)\frac{\Gamma(m + \phi(1-\mu))}{\Gamma(m+1)}\frac{\Gamma(m + k + r - 1)}{\Gamma(m + k + \phi)}q_1^{m+k-1}\right)^{-1}.$$

	Estimate	Std. Err.	t value	$Pr(>\|t\|)$
μ	0.272			
q_0	0.641	0.016	38.858	0.000
q_1	0.982	0.002	494.105	0.000
r	0.252	0.021	11.811	0.000
ϕ	2.320	0.594	3.907	0.000

TABLE 1: MLE for θ for panel data on YouTube visits in Germany.

3 Parameter Estimation

Let $\mathbf{k} = \{k_1, \ldots, k_P\}$ be the number of observed events for all P panelist. Each panelist also has socio-economic indicators such as gender, age, and income. These attributes determine their demographic weight $\tilde{w}_i$, which equals the number of people in the entire population that panelist i represents. Finally, let $w_i = \tilde{w}_i \cdot \left(P / \sum_{i=1}^{P} \tilde{w}_i\right)$ be re-scaled weight of panelist i such that $\sum_{i=1}^{P} w_i$ equals sample size P.
We estimate θ using maximum likelihood (MLE), $\widehat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta; \mathbf{x})$, where the log-likelihood

$$\ell(\theta; \mathbf{x}) = \sum_{\{k \mid x_k > 0\}} x_k \cdot \log \mathbb{P}\left(K = k; \theta\right), \tag{10}$$

and $\mathbf{x} = \{x_k \mid k = 0, 1, \ldots, \max(\mathbf{k})\}$, where $x_k = \sum_{\{i \mid k_i = k\}} w_i$ is the total weight of all panelists with k visits.
For deriving closed form expressions of $\mathbb{P}\left(K = k\right) = \sum_{n=0}^{\infty} \mathbb{P}\left(N = n, K = k\right)$ it is simpler to consider $k = 0$ and $k > 0$ separately:

$$\begin{aligned} \mathbb{P}\left(K = 0\right) = q_0 + (1 - q_0) \times \frac{\Gamma(\phi)}{\Gamma(\phi(1-\mu))} \frac{(1-q_1)^r}{\Gamma(r)} \\ \times \sum_{n=0}^{\infty} \frac{\Gamma(n+1+\phi(1-\mu))}{\Gamma(n+1)} \frac{\Gamma(n+r)}{\Gamma(n+1+\phi)} q_1^n, \end{aligned} \tag{11}$$

and for $k > 0$,

$$\begin{aligned} \mathbb{P}\left(K = k\right) =& (1-q_0)(1-q_1)^r \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1-\mu))} \frac{1}{\Gamma(r)} \times \frac{\Gamma(k+\mu\phi)}{\Gamma(k+1)} \\ & \times \sum_{m=0}^{\infty} (m+k) \frac{\Gamma(m+\phi(1-\mu))}{\Gamma(m+1)} \frac{\Gamma(m+k+r-1)}{\Gamma(m+k+\phi)} q_1^{m+k-1}. \end{aligned} \tag{12}$$

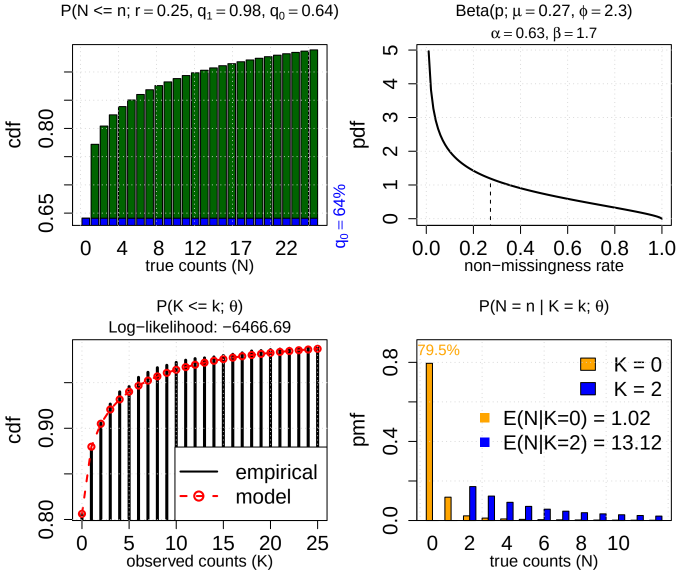

FIGURE 1: Model estimates for: (top left) true counts N_i; (top right) non-missing rate p_i; (bottom left) empirical count frequency and model fit; (bottom right) conditional predictive distributions and expectations.

3.1 Fix expected non-missing rate μ

Usually, researchers must estimate all 5 parameters from panel data. For our application, though, we can estimate (and fix) the non-missing rate μ a-priori as we have access to internal YouTube log files.

Let $\bar{k}_{\tilde{W}} = \sum_{i=1}^{P} \tilde{w}_i k_i$ be the observed panel visits projected to the entire population. Analogously, let $\bar{N}_{\tilde{W}} = \sum_{i=1}^{P} \tilde{w}_i N_i$ be the panel projections of the number of true YouTube visits. While any single N_i is unobservable, we can estimate $\bar{N}_{\tilde{W}}$ by simply counting all YouTube homepage views in Germany from our YouTube log files, yielding $\widehat{\bar{N}}_{\tilde{W}}$. We herewith obtain a plug-in estimate of the non-missing rate, $\widehat{\mu}_{Logs} = \bar{k}_{\tilde{W}} / \widehat{\bar{N}}_{\tilde{W}}$. The remaining 4 parameters, $\theta_{(-\mu)} = (\phi, q_0, r, q_1)$, can be obtained by MLE, $\widehat{\theta}_{(-\mu)} = \arg\max_{\theta_{(-\mu)}} \ell((\widehat{\mu}_{Logs}, \theta_{(-\mu)}); \mathbf{x})$. The overall estimate is $\widehat{\theta} = (\widehat{\mu}_{Logs}, \widehat{\theta}_{(-\mu)})$.

4 Estimating YouTube Audience in Germany

Here we use data from a German online panel (GfK Consumer Panels, 2013), which monitors web usage of $P = 6,545$ individuals in October, 2013 (31 days). In particular, we are interested in the probability that an adult in Germany visited the YouTube homepage `www.youtube.de`. Empirically,

$\widehat{\mathbb{P}}(K = 0) = 0.81$, yielding 19% observed 1+ reach. However, we know by comparison to YouTube log files that the panel only recorded 27.2% of all impressions. We fix the expected non-missing rate at $\widehat{\mu} = 0.272$ and obtain the remaining parameters via MLE (Table 1): Figure 1 shows the model fit for the true, observed, and predictive distribution. In particular, the true 1+ reach is 36% ($\widehat{q}_0 = 0.64$), not 19% as the naïve estimate suggests.

5 Discussion

We introduce a probabilistic framework to impute missing events in count data, including a hurdle component for more flexibility to model lots of zeros. Researchers can use our models to obtain accurate probabilistic predictions of the number of true, unobserved events. We apply our methodology to accurately estimate how many people in Germany visit YouTube.

Acknowledgments: We want to thank Christoph Best, Penny Chu, Tony Fagan, Yijia Feng, Oli Gaymond, Simon Morris, Raimundo Mirisola, Andras Orban, Simon Rowe, Sheethal Shobowale, Yunting Sun, Wiesner Vos, Xiaojing Wang, and Fan Zhang for constructive discussions and feedback.

References

Fader, P. and Hardie, B. (2000). A note on modelling underreported Poisson counts. *Journal of Applied Statistics*, 27(8):953–964.

Ferrari, S. and Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):799–815.

GfK Consumer Panels (2013). Media Efficiency Panel.

Hofler, R. A. and Scrogin, D. (2008). A count data frontier model. Technical report, University of Central Florida.

Hu, M., Pavlicova, M., and Nunes, E. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. *Am J Drug Alcohol Abuse*, 37(5):367–75.

Rose, C., Martin, S., Wannemuehler, K., and Plikaytis, B. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *J Biopharm Stat*, 16(4):463–81.

Schmittlein, D. C., Bemmaor, A. C., and Morrison, D. G. (1985). Why Does the NBD Model Work? Robustness in Representing Product Purchases, Brand Purchases and Imperfectly Recorded Purchases. *Marketing Science*, 4(3):255–266.

Yang, S., Zhao, Y., and Dhar, R. (2010). Modeling the underreporting bias in panel survey data. *Marketing Science*, 29(3):525–539.