How Many Millennials Visit YouTube? Estimating Unobserved Events From Incomplete Panel Data Conditioned on Demographic Covariates

Georg M. Goerg, Yuxue Jin, Nicolas Remy, Jim Koehler

Google Inc.

April 27, 2015

Abstract

Many socio-economic studies rely on panel data as they also provide detailed demographic information about consumers. For example, advertisers use TV and web metering panels to estimate ads effectiveness in selected target demographics. However, panels often record only a fraction of all events due to non-registered devices, technical problems, or work usage. Goerg et al. (4) present a beta-binomial negative-binomial hurdle (BBNBH) model to impute missing events in count data with excess zeros.

In this work, we study empirical properties of the MLE for the BBNBH model, extend it to categorical covariates, introduce a penalized maximum likelihood estimator (MLE) to get accurate estimates by demographic group, and apply the methodology to a German media panel to learn about demographic patterns in the YouTube viewership.

Keywords: imputation; missing data; zero inflation; BBNBH distribution.

1 Introduction

Panels are often used in socio-economic studies to track user activity and estimate characteristics of specific target populations (see 14, for a methodology overview). TV and online media panels (3; 8) are particularly useful for advertisers to estimate the effectiveness of showing an ad at a certain time of the day or placing an ad on a website.

As an example consider Fig. 1a, which shows a random subsample of the panel data we use in our case study in Section 6. The white/black colors encode whether the panel recorded a YouTube homepage visit for a given day from a panelist or not. The most striking feature of the data are a vast amount of zeros, i.e., many panelists seem to never visit `www.youtube.de`. Secondly, the split by gender and age hints at heterogeneity across the population. Figure 1b shows the aggregated

empirical distribution of number of visits – again split by age and gender. Here the difference between demographic groups becomes more pronounced.

Advertisers are interested in *reach* and *frequency*: while the first measures the fraction of the population that sees an ad, the second measures how often they are exposed to it (on average). Reach and frequency can largely determine the cost of an advertising opportunity – on TV, in a magazine, or on a website. It is thus important to obtain accurate and precise reach and frequency estimates from panel data.

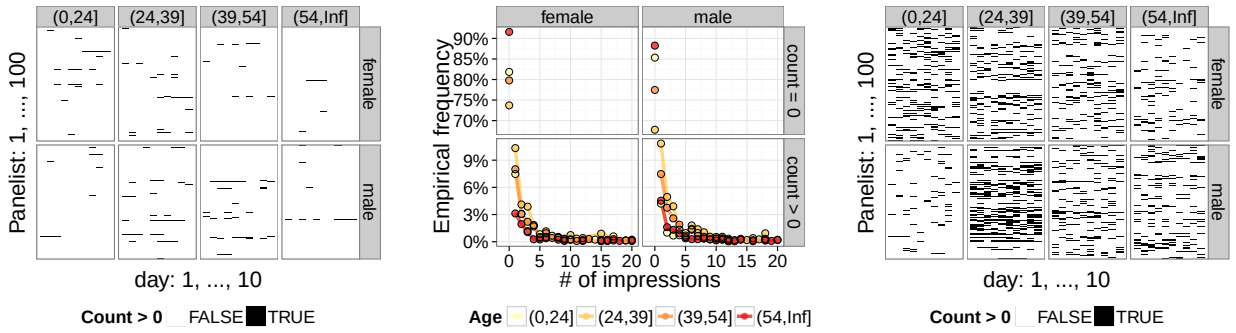
A naïve approach would simply use the sample fraction of positive number of events (website visits, TV spots watched, etc.) to estimate reach; similarly, for a frequency estimate. The problem with this approach is that panels often suffer from underreporting, i.e., they record only a fraction of all events. Missingness can have various causes such as non-compliance, work usage, or the use of unregistered devices (see 11; 12, for a detailed review on the accuracy of panels). This is especially problematic for reach and frequency measurements as missed events always lead to underestimation when using sample averages.

Several studies on response error in consumer surveys have approached the missingness problem. Yang et al. (15) use econometric time series models to estimate response rate over time conditioned on demographic information; Fader and Hardie (1) use a Poisson model with underreporting; Schmittlein et al. (10) derive closed form expressions for predictive distributions of the beta-binomial negative binomial (BBNB) model (6; 5).

Goerg et al. (4) extend the BBNB model with a hurdle component (BBNBH) to account for excess zeros in the data-generating process of the unobserved counts. They derive marginal and predictive distributions and use the methodology to estimate how many people go to the YouTube homepage in Germany. While the BBNBH model can adapt to excess zeros in underlying, true events, it does not take heterogeneity across the population into account. Yet, as advertisers are interested in specific target demographics it is important to get accurate estimates by demographic.

We extend previous work on the BBNBH model (Section 2) in several important ways: a) we add categorical covariates to the BBNBH model and propose a categorical missingness estimation via a penalized maximum likelihood estimator (MLE) in order to capture heterogeneity across categories (e.g., demographic groups or weekend vs. weekday effects) (Section 3); b) we study empirical properties of the MLE via simulations (Section 4); c) we present several methodologies to estimate reach from panel data and the model fits (Section 5); d) we apply the methodology to a German online media panel to estimate demographics differences in the YouTube viewing behavior and provide demographic-specific reach and frequency estimates (Section 6). A preview of these results is shown in Fig. 1c with a random sample of the true (unobserved) non-zero visits as predicted by the demospecific BBNBH model.

Appendix A describes details on estimation algorithms. All computations and figures were done in R (9).



(a) Zero vs. non-zero events of 100 randomly selected panelists in each demographic group during 10 (randomly picked) consecutive days.

(b) Empirical frequency of the number of visits split by gender and age group.

(c) Zero vs. non-zero imputed counts, where parameters are from the overall BBNBH MLE fit (Table 1, Section 6).

Figure 1: Panel data for visiting the YouTube homepage for Germany (`www.youtube.de`). See Section 6 for details.

2 Review of the BBNBH Model: Hierarchical Imputation Of Underreported Count Data

In this section we review the BBNBH model and its main properties.[1] Figure 2 illustrates its data-generating process on the example of YouTube homepage visits:

a) User i visits YouTube with probability $1 - q_0$; if she does, the number of visits N_i per time unit is distributed according to a shifted Poisson distribution (starting at $n = 1$) with rate λ_i.

b) To allow heterogeneity among the population, $\lambda_i \sim \text{Gamma}\left(r, \frac{q_1}{1-q_1}\right)$, with rate $r > 0$ and success probability $q_1 \in (0, 1)$.

c) Given a visit, the probability of recording this visit in the panel equals $p_i \in (0, 1)$.

d) The recording probability p_i follows a Beta distribution, $p_i \sim \text{Beta}(\mu, \phi)$, with an expected non-missing rate μ and precision ϕ, which characterizes the variability of missingness across the population.[2]

Combining a) and b) yields a negative binomial hurdle (NBH) distribution for the unobserved events $N_i \geq 0$ with probability mass function (pmf)

$$NBH(n; q_0, q_1, r) = \begin{cases} q_0, & \text{if } n = 0, \\ (1 - q_0) \cdot \frac{\Gamma(n+r-1)}{\Gamma(r)\Gamma(n)} \cdot (1 - q_1)^r q_1^{n-1}, & \text{if } n \geq 1, \end{cases} \tag{1}$$

[1] For detailed derivations see Goerg et al. (4).

[2] Mean μ and precision ϕ are related to the (α, β) parametrization of the Beta distribution by $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \alpha + \beta$, respectively; vice versa, $\alpha = \phi\mu$ and $\beta = \phi(1 - \mu)$ (2).

where $\Gamma(z)$ is the gamma function. Levels c) and d) describe a Beta-binomial (BB) subsampling to obtain observed events $K_i \geq 0$ with pmf:

$$BB(k \mid n; \mu, \phi) = \binom{n}{k} \frac{B(k + \phi\mu, n - k + \phi(1-\mu))}{B(\phi\mu, \phi(1-\mu))}, \tag{2}$$

where $B(a, b)$ is the Beta function. Jointly, they constitute the BBNBH model hierarchical imputation

$$\begin{aligned} N_i &\sim NBH(N; q_0, r, q_1), \\ K_i \mid N_i &\sim BB(K \mid N_i; \mu, \phi), \end{aligned} \tag{3}$$

with parameter vector $\theta = (\mu, \phi, q_0, r, q_1)$.

The hurdle parameter q_0 plays an important role in the advertising context as $1 - q_0$ equals the true, but unobserved, potential 1+ reach of a campaign (see Section 5 for details). That is, if an advertiser shows an ad on the YouTube homepage they can expect that a fraction of $1 - q_0$ of the population sees the ad at least once in a given period.

Before deriving marginal and predictive distributions, consider the expected number of true and observed events. It holds,

$$\begin{aligned} \mathbb{E}N &= \mathbb{E}(N \mid N = 0) \cdot \mathbb{P}(N = 0) + \mathbb{E}(N \mid N > 0) \cdot \mathbb{P}(N > 0) \\ &= (1 - q_0) \cdot \left(1 + r\frac{q_1}{(1 - q_1)}\right), \end{aligned} \tag{4}$$

and by the law of total expectation,

$$\mathbb{E}K = \mu \cdot \mathbb{E}N = \mu \cdot (1 - q_0) \cdot \left(1 + r\frac{q_1}{(1 - q_1)}\right). \tag{5}$$

While the analytical derivations of marginal and predictive distributions are simpler for the (r, q_1) parametrization of the negative-binomial, in applications it is useful to consider the mean in (4) as an intuitive and directly interpretable quantity of the model.

2.1 Marginal and predictive distributions

The marginal distribution of the observable events K equals

$$\begin{aligned} \mathbb{P}(K = 0) = q_0 + (1 - q_0) \times &\frac{\Gamma(\phi)}{\Gamma(\phi(1-\mu))} \frac{(1 - q_1)^r}{\Gamma(r)} \\ &\times \sum_{n=0}^{\infty} \frac{\Gamma(n + 1 + \phi(1-\mu))}{\Gamma(n+1)} \frac{\Gamma(n + r)}{\Gamma(n + 1 + \phi)} q_1^n, \end{aligned} \tag{6}$$

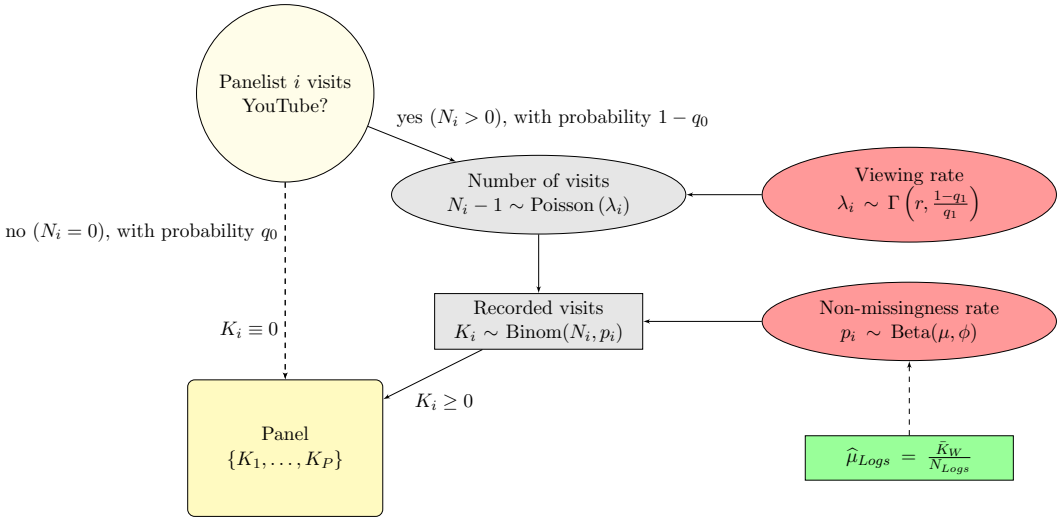


Figure 2: Data-generating process of the BBNBH model for observed – but underreported – count data, illustrated on the example of YouTube homepage visits recorded in a panel.

and for $k > 0$,

$$\begin{aligned}\mathbb{P}(K = k) =& (1-q_0)(1-q_1)^r \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1-\mu))} \frac{1}{\Gamma(r)} \times \frac{\Gamma(k+\mu\phi)}{\Gamma(k+1)} \\ & \times \sum_{m=0}^{\infty} (m+k) \frac{\Gamma(m+\phi(1-\mu))}{\Gamma(m+1)} \frac{\Gamma(m+k+r-1)}{\Gamma(m+k+\phi)} q_1^{m+k-1}.\end{aligned} \tag{7}$$

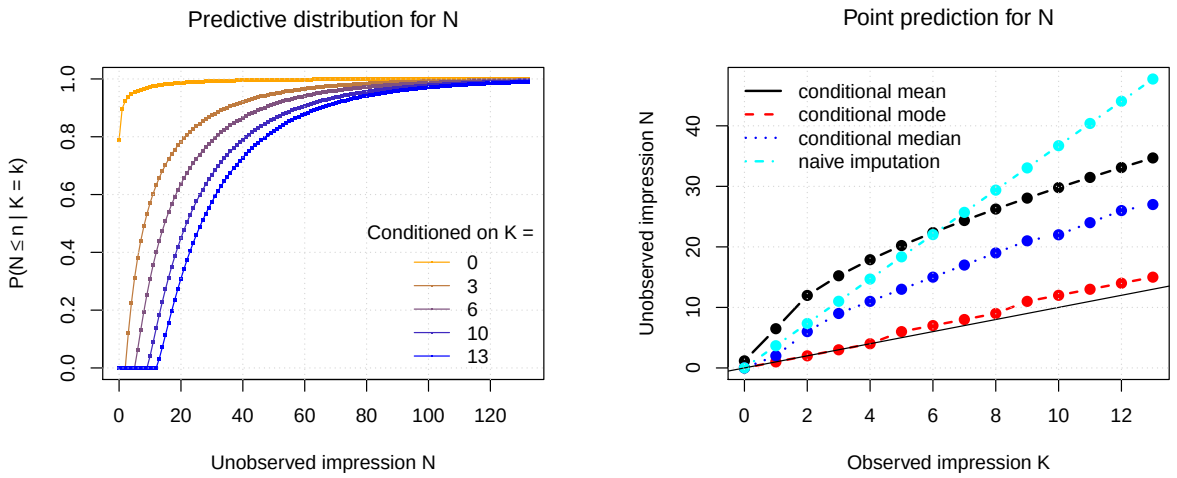
While the panel only records k_i events for each panelist, it is clearly important to find out how many events n_i truly occurred. That is, we are interested in the conditional distribution $\mathbb{P}(N = n_i \mid K = k_i)$. Following Bayes' rule (dropping subscripts) this can be expressed as

$$\mathbb{P}(N = n \mid K = 0) = \frac{1}{\mathbb{P}(K = 0)} \cdot \begin{cases} q_0, & \text{if } n = 0, \\ \frac{\Gamma(n+\phi(1-\mu))}{\Gamma(n+\phi)} \frac{\Gamma(\phi)}{\Gamma(\phi(1-\mu))} & \\ \times (1-q_0) \frac{\Gamma(n+r-1)}{\Gamma(n)} \frac{(1-q_1)^r}{\Gamma(r)} q_1^{n-1}, & \text{otherwise.} \end{cases} \tag{8}$$

and

$$\mathbb{P}(N = n \mid K = k) = \begin{cases} 0, & \text{for all } n < k, \\ n \cdot q_1^{n-1} \frac{\Gamma(n-k+(1-\mu)\phi)}{\Gamma(n-k+1)\Gamma(n+\phi)} \Gamma(n+r-1) & \\ \times \left(\sum_{m=0}^{\infty} (m+k) \frac{\Gamma(m+\phi(1-\mu))}{\Gamma(m+1)} \frac{\Gamma(m+k+r-1)}{\Gamma(m+k+\phi)} q_1^{m+k-1} \right)^{-1}, & \text{otherwise .} \end{cases} \tag{9}$$

(10)



(a) Conditional cdf

(b) Point predictions: Observed impressions vs. mean, median, mode, and naïve imputation ($\widehat{n} = k \cdot 1/\widehat{\mu}_{Logs}$)

Figure 3: Conditional inference via imputation; parameters from Section 6, Table 1.

Figure 3 shows the conditional cumulative distribution functions (cdf) for several k_i and a comparison of several point predictions (expectation, median, mode, and naïve imputation).

3 Parameter Estimation

In practice, the parameter vector $\theta = (\mu, \phi, q_0, r, q_1)$ must be estimated from the panel. Let $\mathbf{k} = \{k_1, \ldots, k_P\}$ be the number of observed events for each panelist $i = 1, \ldots, P$.

Panelists are usually also associated with demographic and economic indicators such as gender, age, and income. Based on these attributes a panelist i has a demographic weight $\tilde{w}_i$ that equals the number of people they represent in the population. A representative panel should be designed such that the total panel weight, $\tilde{W} = \sum_{i=1}^{P} \tilde{w}_i$, equals the total population count (obtained from, e.g., census data). Finally, let $w_i = \frac{\tilde{w}_i}{\tilde{W}} \cdot P$ be the re-scaled weight of panelist i such that the sum of all weights equals the sample size P.

The likelihood of θ

$$\ell(\theta; \mathbf{x}) = \sum_{\{k | x_k > 0\}} x_k \cdot \log \mathbb{P}\left(K = k; \theta\right). \tag{11}$$

depends on the sufficient statistic, $\mathbf{x} = \{x_k \mid k = 0, 1, \ldots, \max(\mathbf{k})\}$, where $x_k = \sum_{\{i | k_i = k\}} w_i$ is the total weight of panelists with k visits. That is, $\mathbf{x}$ is a weighted frequency table of panel counts.

The maximum likelihood estimator (MLE)

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta; \mathbf{x}) \tag{12}$$

can be obtained by numerical optimization. For a covariance matrix (and standard error) estimate of $\widehat{\theta}$ we use the inverse of the numerically obtained Hessian at the optimum.

3.1 Heavy tail robustness: right-truncated log-likelihood

We have found empirically that panel observations do not only have excess zeros, but also heavy tails. That is, some panelists have an extremely high number of recorded visits. Figure 4 shows the overall ecdf of the panel with some extremely large counts. To make the estimation more robust to these extremes, we right-truncate the summation in the log-likelihood at some $k = k_q$, and then add the cumulative probability for the event $\{K > k_q\}$.

Formally, we approximate the exact log-likelihood in (11) with

$$\ell(\theta; \mathbf{x})_{trunc} = \sum_{\{k | x_k > 0, k \leq k_q\}} x_k \cdot \log \mathbb{P}\left(K = k; \theta\right) + \left(\sum_{k > k_q} x_k \right) \cdot \log \mathbb{P}\left(K > k_q; \theta\right). \tag{13}$$

Since K grows with the length of the time period (events per day, week, month, etc.), it is not possible to propose a generally good truncation value. We thus choose k_q based on the empirical quantile q as it adapts automatically to the time scale. We found that using $q \approx 0.99$ works well in practice.

3.2 Fix expected non-missing rate μ

The optimization in (12) takes place over a 5-dimensional parameter space, $(\mu, \phi, q_0, r, q_1) \in \Theta = (0,1) \times \mathbb{R}^+ \times (0,1) \times \mathbb{R}^+ \times (0,1)$. As we have access to internal YouTube log files we can reduce it to 4 dimensions as we can fix the expected non-missing rate μ a-priori by comparing panel data with log files.

Let $\bar{k}_{\tilde{W}} = \sum_{i=1}^{P} \tilde{w}_i k_i$ be the observed visits projected to the entire population. Analogously, let $\bar{N}_{\tilde{W}} = \sum_{i=1}^{P} \tilde{w}_i N_i$ be the panel count of the number of true homepage visits of the entire population. While each single N_i is unobservable, we know what $\bar{N}_{\tilde{W}}$ should be by counting all visits to YouTube from our internal log files. This estimate, $\widehat{\bar{N}}_{\tilde{W}}$, can be used to get a fixed plug-in estimate of the expected non-missing rate, $\widehat{\mu}_{Logs} = \bar{k}_{\tilde{W}} / \widehat{\bar{N}}_{\tilde{W}}$. The remaining four parameters, $\theta_{(-\mu)} = (\phi, q_0, r, q_1)$, can be obtained by MLE:

$$\widehat{\theta}_{(-\mu)} = \arg\max_{\theta_{(-\mu)}} \ell((\widehat{\mu}_{Logs}, \theta_{(-\mu)}); \mathbf{x}). \tag{14}$$

The overall estimate is $\widehat{\theta} = (\widehat{\mu}_{Logs}, \widehat{\theta}_{(-\mu)})$.

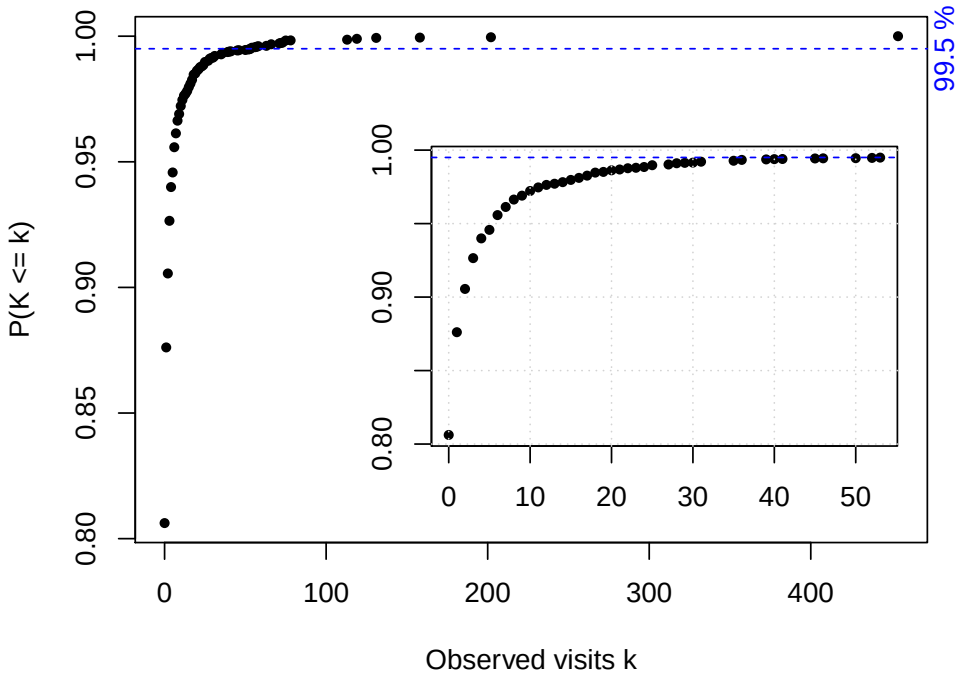


Figure 4: Empirical cdf (weighted) of panel counts. The horizontal blue line shows the truncation at the $q = 99.5$ % quantile. The sub-plot shows the same ecdf, but with the x-axis restricted to counts below this quantile, $k = 0, \ldots, 54$.

In simulations and applications we found that fixing μ gives much more stable estimates, especially with respect to q_0 and r. See also Section 4.

3.3 Demographic-dependent estimation

Advertisers use panels to measure viewing behavior of specific target audiences, e.g., young females. Figure 1b shows that panel observations vary strongly across demographic groups. The basic BBNBH model and resulting reach and frequency estimates in Goerg et al. (4), however, rely on the same $\widehat{\theta}$ for all panelists and hence do not provide good demographic-specific inference.

We thus extend the BBNBH model with categorical covariates thus having category-dependent parameters, $\theta^{(1:G)} = \left(\theta^{(1)}, \ldots, \theta^{(G)}\right)$ – one for each of G exhaustive sub-groups of panel observations, $D^{(1)}, \ldots, D^{(G)}$ (e.g., $D^{(1)}$ = "female", $D^{(2)}$ = "male" or $D^{(1:7)} = \{\text{Mon}, \ldots, \text{Sun}\}$). Such a model has $5 \times G$ parameters. For the remainder of this work we will use demographic groups as the categorical covariate. Note though, that the methodology carries over to other categories such as weekday effects or economic status.

Conditioning on demographic sub-groups, the log-likelihood becomes

$$\ell^{(G)}(\theta^{(1:G)}; \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(G)}) = \sum_{g=1}^{G} \ell(\theta^{(g)}; \mathbf{x}^{(g)}), \tag{15}$$

where $\mathbf{x}^{(g)} = \{x_k^{(g)} \mid k = 0, 1, 2, \ldots\}$ and $x_k^{(g)}$ is the total weight of panelists in $D^{(g)}$ with k views. Since splitting by demographic yields non-overlapping subgroups of panelists with independent parameters, (15) can be maximized for each subgroup separately; thus $\widehat{\theta}^{(1:G)} = \left(\widehat{\theta}^{(1)}, \ldots, \widehat{\theta}^{(G)}\right)$.

3.3.1 Fix expected non-missing rate per demographic

As for the overall model, fixing non-missing rates greatly improves estimation stability. For multiple groups there are at least three ways to fix $\mu^{(1:G)} = \left(\mu^{(1)}, \ldots, \mu^{(G)}\right)$:

Plugin demo-specific missingness: Missing rates can be estimated for each subgroup separately (comparing panel vs. log files to obtain $\widehat{\mu}_{Logs}^{(g)}$), and then one proceeds as with the overall MLE.

While this approach is appealing for its simplicity, it suffers from estimation bias since online demographic information is often incomplete and not entirely correct: younger users tend to report an older age online and desktop devices are often shared between family members in a household. Thus comparing (correct) panel demographic information to logs data yields biased non-missing rate estimates.[3]

All equal missingness: One option to avoid this particular estimation bias, is to set $\mu^{(g)} \equiv \widehat{\mu}_{Logs}$ for all $g = 1, \ldots, G$.

While this conveniently reduces the parameter space to $4 \times G$ dimensions, missing rates do not depend on demographics anymore. However, it is quite unrealistic since causes for missingness, such as home vs. work usage or using multiple devices, are strongly correlated with demographics.

Variable, but overall fixed, missingness: Our suggested approach allows missing rates $\mu^{(1:G)}$ to vary by demographic group, but we restrict them to average out to the overall $\widehat{\mu}_{Logs}$.

To obtain the constraint on $\mu^{(1:G)}$ we reason as follows. Let $k_{\tilde{W}}^{(g)} = \sum_{i \in D^{(g)}} \tilde{w}_i k_i$ be the total number of observed visits by demographic group $D^{(g)}$. Analogously, let $N_{\tilde{W}}^{(g)}$ be the total number of true (unobserved) events per group. Note again that $k_{\tilde{W}}^{(g)}$ can be computed from the panel, while $N_{\tilde{W}}^{(g)}$ is unobserved.

[3]Hence we do not pursue this approach in our case study at all. Correcting demographic estimation bias of YouTube logs files is beyond the scope of this work (see Wang and Koehler, for details).

By construction the observed events per group add up to the overall number of events:

$$\sum_{g=1}^{G} k_{\tilde{W}}^{(g)} = k_{\tilde{W}}. \tag{16}$$

Clearly, (16) must also hold for the true events,

$$\sum_{g=1}^{G} N_{\tilde{W}}^{(g)} = \bar{N}_{\tilde{W}}. \tag{17}$$

To obtain a restriction on $\mu^{(1:G)}$ note that if we could obtain the total number of true visits for each group without the demographic-estimation bias, then we could estimate demographic non-missing rates for each group using $\widehat{\mu}^{(g)} = \frac{k^{(g)}}{N^{(g)}}$. Thus, (17) can be rewritten as

$$\sum_{g=1}^{G} \frac{k^{(g)}}{\widehat{\mu}^{(g)}} = \frac{k_{\tilde{W}}}{\widehat{\mu}_{Logs}} \Leftrightarrow \frac{1}{\sum_{g=1}^{G} v^{(g)} \frac{1}{\widehat{\mu}^{(g)}}} = \widehat{\mu}_{Logs}, \tag{18}$$

where $v^{(g)} = \frac{k_{\tilde{W}}^{(g)}}{k_{\tilde{W}}}$ are data-driven weights. Constraint (18) states that the weighted *harmonic* average of groupwise non-missing rate estimates must equal the overall non-missing rate. For a fixed $\widehat{\mu}_{Logs} > 0$, (18) avoids degenerate $\mu^{(g)} \to 0$ optima. It is important to point out that the non-missing rates are not weighted by demographic weights, but by the (weighted) number of counts in each group.

3.3.2 Iterative exact-constraint estimator

As (18) binds parameters from different groups together, the MLE cannot be solved separately for each group. jointly subject to (18). However, while non-missing rates $\mu^{(1:G)}$ are tied by (18), remaining parameters do not influence each other across groups. It is therefore not necessary to perform joint maximization in the entire $5 \times G - 1$ dimensional space, but optimization can be done iteratively to accelerate convergence:

0. Use overall $\widehat{\theta}$ as starting value for each group: $\widehat{\theta}_0^{(g)} = \widehat{\theta}$. Set $i = 1$.

1. For each $g \in \{1, \ldots, G\}$: fix $\widehat{\mu}_{i-1}^{(g)}$ and solve (14) to obtain $\widehat{\theta}_{(-\mu),i}$.

2. Fix $\widehat{\theta}_{(-\mu),i}^{(g)}$ of each group and maximize log-likelihood in (15) over $\mu^{(1:G)}$ subject to (18) to obtain $\widehat{\mu}_i^{(1:G)}$. Set $i = i + 1$.

3. Iterate steps 1 and 2 until convergence, $\|\widehat{\theta}_{i-1}^{(1:G)} - \widehat{\theta}_i^{(1:G)}\| < \varepsilon$, for some tolerance level $\varepsilon > 0$.

Since step 1 is an unconstrained optimization, the MLE from (14) can be used. Step 2 requires solving a constrained optimization problem. To get an exact solution we map the G dimensional

$\mu^{(1:G)}$ subject to constraint (18) to the unbounded $\mathbb{R}^{G-1}$, optimize on the unconstrained space, and then map it back to the original space. This bijective mapping guarantees that $\widehat{\mu}_i^{(1:G)}$ satisfies (18) exactly in each iteration. For details see Appendix A.2.

In practice, the iterative updating between $\mu^{(1:G)}$ and remaining parameters can lead to a "zig-zagging" of the estimates. To make these transitions more smooth in each iteration we use a weighted average between old $(i-1)$ and new (i) estimates of the form (similarly for $\widehat{\theta}_{i,-\mu}^{(1:G)}$)

$$\widehat{\mu}_i^{(1:G)} \leftarrow \lambda \cdot \widehat{\mu}_i^{(1:G)} + (1-\lambda) \cdot \widehat{\mu}_{i-1}^{(1:G)}, \tag{19}$$

where $\lambda \in (0,1]$ controls the smoothness. For $\lambda = 1$ no smoothing occurs; for $\lambda \to 0$ the transitions become more smooth.

3.4 Smoothing penalty on variation of missingness

To avoid too large variation across $(\mu_1, \ldots, \mu_G)$ we add a penalty term to the log-likelihood

$$\kappa \cdot \| \operatorname{dist}\left(\left(\widehat{\mu}_{Logs}, \ldots, \widehat{\mu}_{Logs}\right), \mu^{(1:G)} \right) - \delta \|_1 \tag{20}$$

where $\operatorname{dist}(\cdot,\cdot)$ is a distance measure, $\|x\|_1 = |x|$ is the absolute value of x, $\delta \geq 0$ is the expected target distance, and the penalty parameter $\kappa \geq 0$ regulates the deviation from δ. If $\mu^{(1)} = \ldots = \mu^{(G)} = \widehat{\mu}_{Logs}$ then (20) equals δ. From a Bayesian point of view, $\delta > 0$ encodes the belief that missing rates are not expected to be constant ($\delta = 0$), but should vary across groups by a variation of δ (as measured by $\operatorname{dist}(\cdot,\cdot)$). Thus while we set a target variation a-priori, we let the data (likelihood) decide which groups are below and which ones are above the overall missingness.

For the distance measure we use a weighted L_p norm, $\operatorname{dist}(x,y) = \|x-y\|_{p,v} = \left(\sum_{g=1}^{G} v^{(g)} |x_g - y_g|^p\right)^{1/p}$, $p = 2$, where the (re-scaled) demographic weights per group, $v^{(g)} = \frac{w^{(g)}}{W} G$, satisfy $v^{(g)} \geq 0$, $\sum_{g=1}^{G} v_g = G$.

4 Simulations

This section presents empirical finite-sample size properties of the MLE with particular focus on the performance improvements when fixing $\widehat{\mu}$ a-priori.

We simulate panels of size $P \in \{100, 500, 1000, 2000, 5000, 10000\}$ ($w_i = 1$ for all i). We use typical parameters found in our case study, $\theta = (\mu = 0.25, \phi = 3, q_0 = 0.8, r = 0.5, q_1 = 0.95)$: this means that 20% of the population visit at least once; those who visit typically see 10.5 impressions ($\mathbb{E}(N \mid N > 0) = 1 + r\frac{q_1}{1-q_1}$); but – on average – only 25% of their visits are recorded in the panel. We have found via simulations and several applications that the log-likelihood surface is very

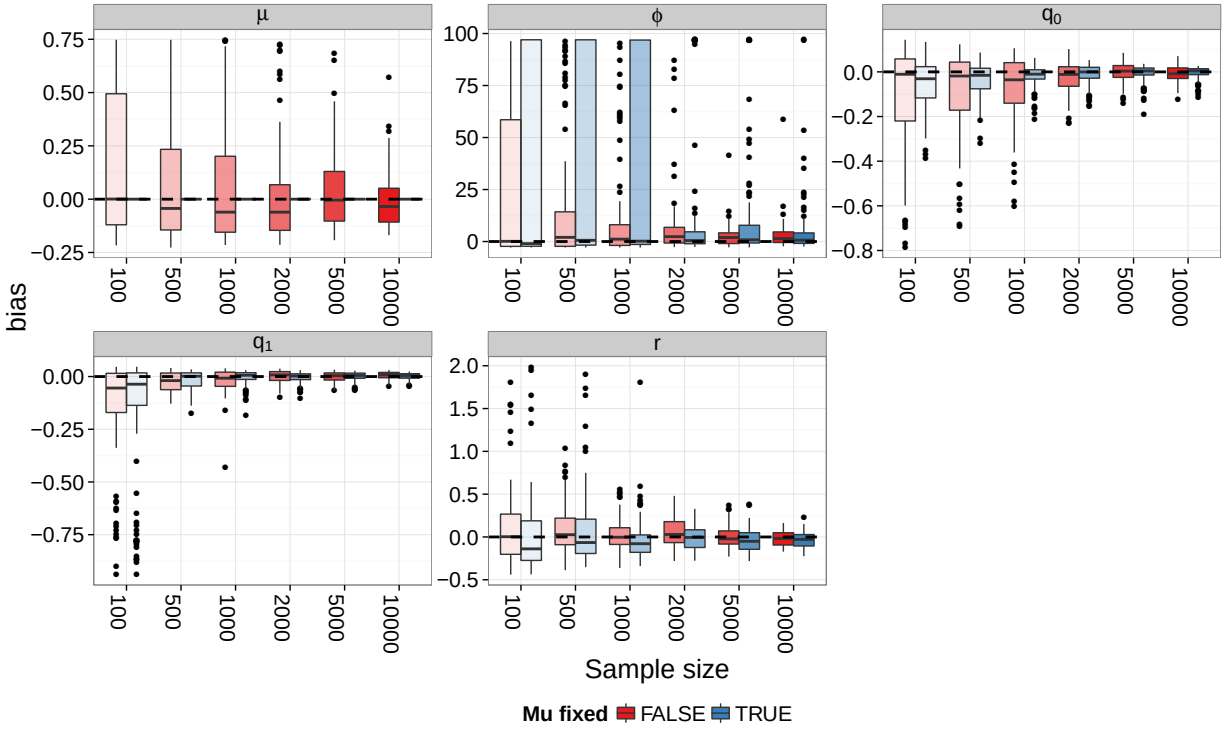


Figure 5: Estimation bias as a function of sample size and for two types of estimators: estimating μ by MLE (red) versus fixing $\widehat{\mu}$ a priori (blue). As $\widehat{r}$ can be highly variable for $P = 100$ we truncate the y-axis for r at the 80% quantile (bias ≥ 2.1) for better readability.

unstable in the (q_0, μ) sub-space. A pure gradient method often fails to estimate all 5 parameters jointly as it drifts off to local optima (often $\widehat{q_0} \to 0$ and $\widehat{\mu} \to 0$).[4] We thus suggest to use a differential evolution algorithm to maximize (11) over all 5 parameters. In the simulations we use the `DEoptim` package in R (9). While the random search component in `DEoptim` is more robust to local optima, it takes longer to converge. It is therefore equally important to provide good starting value θ_0. See Appendix A.1.1 for details.

Results from $n = 100$ replications in Figure 5 show that the MLE is accurate even for small P, but has large variance – especially for the two most important parameters: non-missing rate (μ) and 1+ reach ($1 - q_0$). This large uncertainty can be reduced by using prior information on $\widehat{\mu}$ (setting $\widehat{\mu} = \mu$) – a consequence of reducing the parameter space from 5 to 4 dimensions. In particular, $\widehat{q_0}$ is much more reliable.

As a function of sample size the MLE behaves as expected in most cases: variability decreases as P increases. Interestingly though, $\widehat{\phi}$ is considerably worse for small P when μ is fixed a-priori; only for $P \geq 2,000$ estimates of ϕ have lower variability. Remaining parameters, on the other hand, have the expected lower variablity when fixing $\mu = \widehat{\mu}$; especially for $P \geq 1,000$.

[4] The model would thus tell us that everybody visits YouTube ($1 - q_0 = 1$), but the panel misses everything ($\mu = 0$).

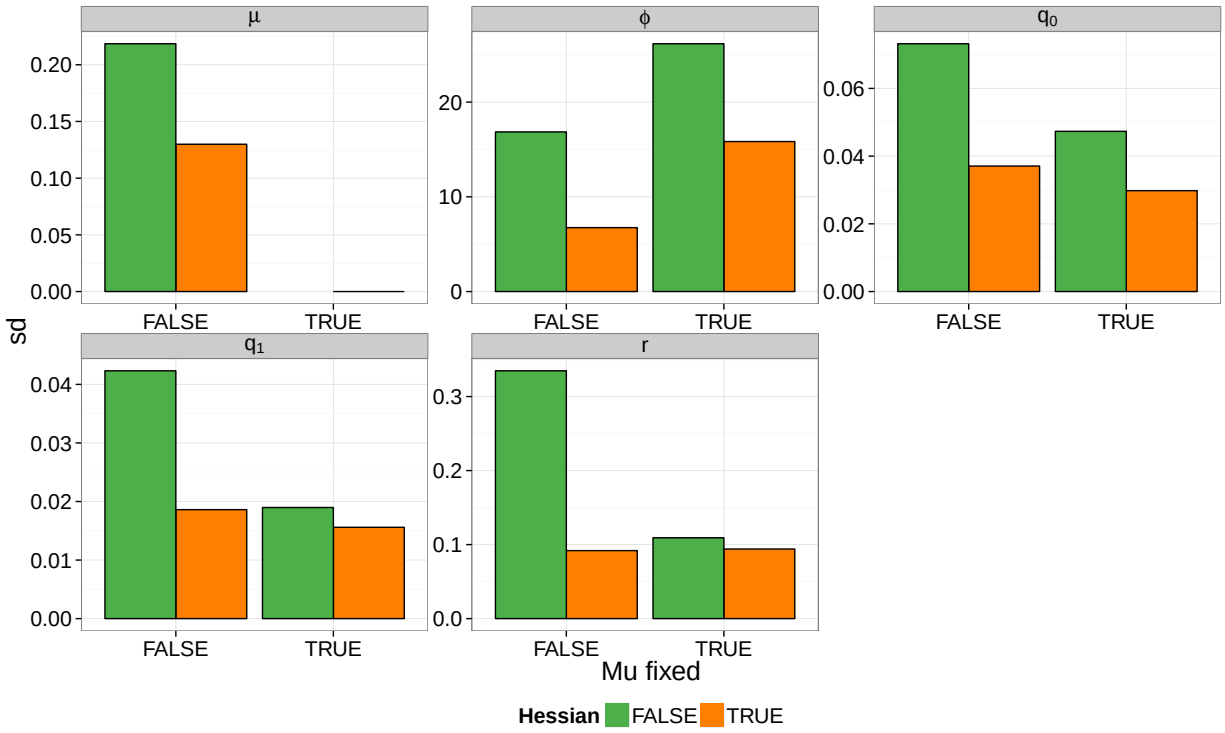


Figure 6: Standard error comparison for sample size $P = 10,000$: empirical standard deviations of $\widehat{\theta}$ (green) versus average of standard errors determined by inverse of the numerical Hessian (orange).

Thus while our proposed methodology can estimate non-missing rate and 1+ reach accurately from the panel alone (unbiased), it has poor precision. Only with a good prior estimate of μ and large enough sample size can good precision be achieved.[5]

Figure 6 compares the numerical standard error estimates (average over 100 replications of $se(\widehat{\theta}_i)$ for each run) to the sample standard deviation of the $n = 100$ estimates ($\widehat{\sigma}\left(\widehat{\theta}_i\right)$).[6] It shows that standard errors obtained by diagonal of the inverse Hessian (orange) underestimate the sampling variation (green) – and thus lead to too narrow confidence intervals. When fixing $\widehat{\mu} = \mu$ a-priori the difference becomes much smaller.

Overall, simulations show that our proposed model is identifiable and researchers can use maximum likelihood to obtain unbiased parameter estimates. However, standard errors from the Hessian are much smaller than expected under repeated sampling. We thus suggest boot-strapping to get confidence intervals with proper coverage probability when estimating μ from the data as well.

[5] Future work can extend this to a prior distribution on μ rather than a point estimate $\widehat{\mu}_{Logs}$.

[6] We only show estimates for sample size $P = 10,000$. However, as expected, the difference between theoretical and empirical standard errors get larger as P decreases.

5 Reach Estimation

One main objective of monitoring the panel is to estimate the $\ell+$ reach of a historical campaign on the YouTube homepage. That is, advertisers want to know the fraction of the target population that has seen their ad at least ℓ times. Formally, $\ell+$ reach can be computed as

$$R_D(\ell) = \frac{1}{W_D} \sum_{i \in D} w_i \cdot \mathbb{P}\left(\text{panelist } i \text{ visited at least } \ell \text{ times}\right), \tag{21}$$

where $D \subseteq \{1, \ldots, P\}$ can be a subset of panelists representing the target population, e.g., males between 18 and 34 years old, and $W_D = \sum_{i \in D} w_i$ is the total demographic weight of D. Note that the demographic categories from the estimation do not necessarily have to match the target demographic D of an advertiser. We will thus below sum over all categories from the estimation.

If panels would not suffer from underreporting ($\mu = 1$), then one could estimate (21) using a sample average

$$\widehat{R}_D(\ell)_{empirical} = \frac{1}{W_D} \sum_{i \in D} w_i \cdot \mathbb{1}(k_i \geq \ell). \tag{22}$$

A disadvantage of (22) is that the indicator function leads to noisy estimates (especially for large ℓ) and $\mathbb{1}(k_i \geq \ell) = 0$ for $\ell > \max(\mathbf{k})$. To obtain smooth and non-zero estimates for large ℓ we suggest to use the fitted marginal distribution of K in (21)

$$\widehat{R}_D(\ell)_{observable} = \frac{1}{W_D} \sum_{g=1}^{G} \left(\sum_{i \in D^{(g)}} w_i \cdot \mathbb{P}\left(K \geq \ell; \widehat{\theta}^{(g)}\right) \right). \tag{23}$$

Similarly to data vs. model fit checks, differences between (22) and (23) indicate a poor model fit.

In practice, though, $\mu < 1$ and (22) & (23) underestimate historical $\ell+$ reach. For an unbiased estimate we thus recommend to use the conditional probability that panelist i has visited at least ℓ times *given* the panel recorded k_i visits

$$\widehat{R}_D(\ell)_{imputed} = \frac{1}{W_D} \sum_{g=1}^{G} \left(\sum_{i \in D^{(g)}} w_i \cdot \mathbb{P}\left(N \geq \ell \mid K = k_i; \widehat{\theta}^{(g)}\right) \right). \tag{24}$$

In Section 6 we demonstrate that (24) can be significantly larger than (22).

For the sake of completeness we also present the unconditional, true $\ell+$ reach estimate

$$\widehat{R}_D(\ell)_{unobservable} = \frac{1}{W_D} \sum_{g=1}^{G} \left(\sum_{i \in D^{(g)}} w_i \cdot \mathbb{P}\left(N \geq \ell; \widehat{\theta}^{(g)}\right) \right). \tag{25}$$

The difference between (24) and (25) is that $\widehat{R}_D(\ell)_{imputed}$ estimates the *historical* unobserved reach, whereas $\widehat{R}_D(\ell)_{unobservable}$ estimates it for (another) *potential* realization of the panel (assuming

	Estimate	Std. Err.	t value	$Pr(>\|t\|)$
μ	0.272			
q_0	0.636	0.023	27.990	0.000
q_1	0.976	0.004	277.949	0.000
r	0.298	0.030	9.925	0.000
ϕ	1.941	0.644	3.015	0.003

Table 1: MLE given a-priori fixed $\mu = \widehat{\mu}_{Logs}$ with the truncated log-likelihood in (13) with $k_q = 54$.

stationarity over time).

6 Case Study: Imputing YouTube Homepage Impressions

We now illustrate the imputation methodology on data from an online panel in Germany, which monitors YouTube usage for the period from 2013-10-01 to 2013-10-31 (31 days). After data-cleaning, we remain with $6,545$ panelists representing the adult online population of Germany. For this analysis, we focus on estimating $\ell+$ reach of the YouTube homepage (`www.youtube.de`) from desktop devices only.[7]

Figure 4 shows the empirical cumulative distribution function (ecdf) of the panel, where counts k_i have been weighted by the demographic weight w_i of panelist i. Even over a period of 31 days, the proportion of zero visits is quite high ($\widehat{\mathbb{P}}(K = 0) = 0.81$). On the other extreme, the panel also shows several outliers ($\max(k_i) = 454$), which make our proposed robust right-truncated log-likelihood approach from Section 3.1 worthwhile.

6.1 Parameter estimation

Our internal YouTube log files for Germany show that the panel has a non-missing rate of $\widehat{\mu}_{Logs} = 0.27$. Contrary to simulations, the MLE for all 5 parameters on this dataset suffers from the instability in the (μ, q_0) subspace ($\widehat{\mu} \to 0$ and $\widehat{q}_0 \to 0$; further results not shown).[8] We thus proceed with a fixed non-missing rate and estimate remaining parameters using the truncated log-likelihood approach.

A comparison of the ecdf to the estimated theoretical cdf of K (bottom-left panel of Figure 7) shows that the estimated model (Table 1) provides an excellent fit to data. For model interpretation, first consider the NBH part: the estimated hurdle probability lies at $\widehat{q}_0 = 0.64$, the negative-binomial parameters are $\widehat{r} = 0.3$ and $\widehat{q}_1 = 0.98$. The fitted distribution of true, unobserved counts N_i (top-

[7] Even though mobile devices play a significant role in today's web usage, we want to stress that our case study does *not* include visits from mobile devices.

[8] We are currently working on an extension of impression imputation to a cookie & impression imputation model, which avoids this instability. We refer to future work.

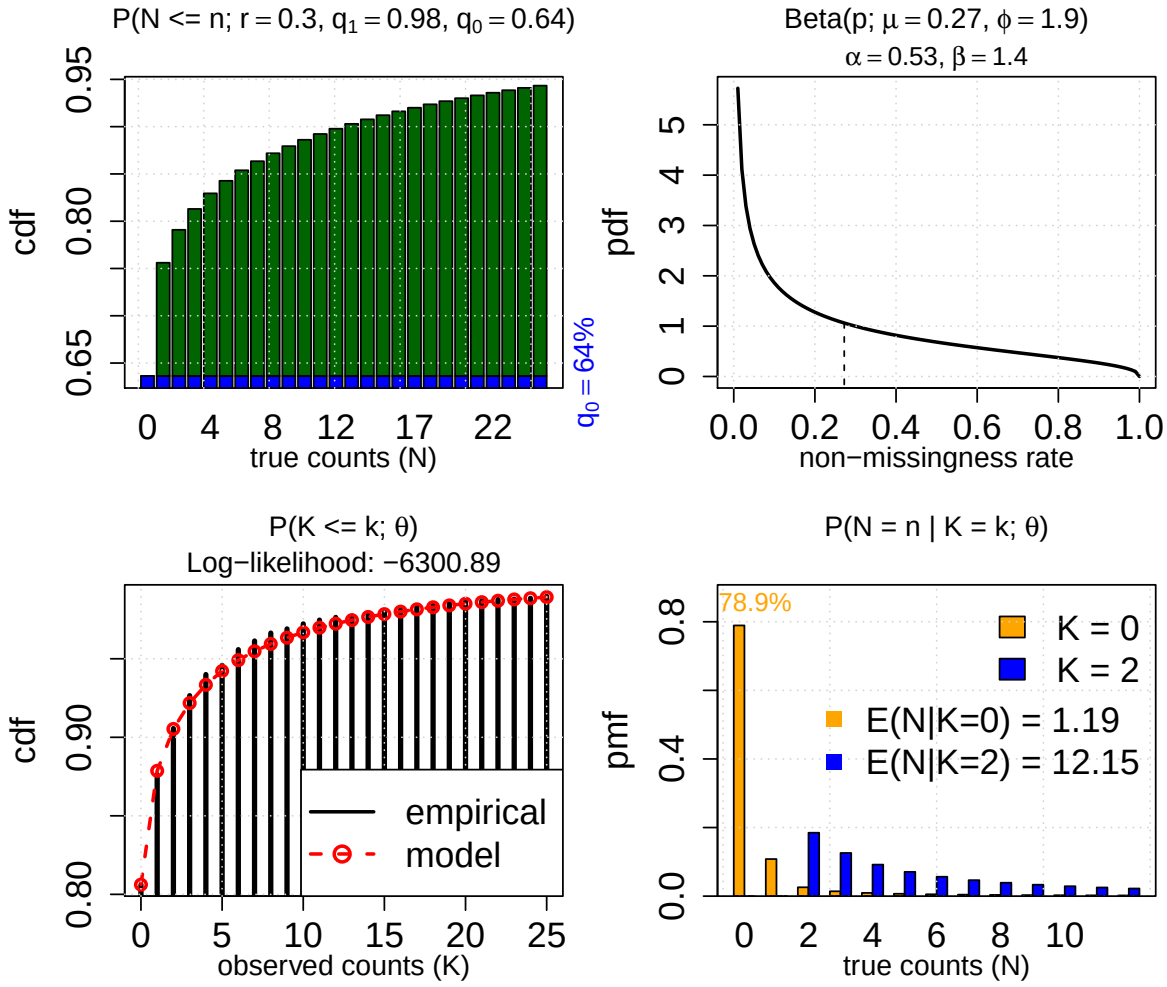


Figure 7: Model check for right-truncated log-likelihood maximization: distribution of true counts N (top left); Beta prior on sub-sampling probability (top right); empirical distribution and model fit (bottom left); conditional predictive distributions for imputation along with conditional expectations (bottom right).

left panel of Fig. 7) shows that the excess zeros in the panel are not solely due to high missingness, but also a consequence of a high probability of not visiting the YouTube homepage at all (63.6%). Secondly, the sub-sampling is characterized by the Beta prior (top right). As $\mu = 0.27$ was fixed, the MLE could adjust its shape via the precision parameter ($\widehat{\phi} = 1.94$): it puts a high mass at a very low non-missing rate; that is, more often than not, almost none of the panel visits were recorded. As a consequence the empirical 1+ reach, $\widehat{\mathbb{P}}(K \geq 1) = 19.4\%$, largely underestimates the true 1+ reach estimate $\mathbb{P}\left(N \geq 1 \mid \widehat{\theta}\right) = 1 - \widehat{q}_0 = 36.4\%$.

6.2 Imputation

The question we are trying to answer is, of course, how many impressions did user i really see *given* she saw k_i impressions. This is particularly important for $k_i = 0$: here it seems that user i has

not visited the YouTube homepage at all, when in fact, he visited ℓ times with positive probability $\mathbb{P}(N_i = \ell \mid k_i = 0) > 0$.

Since $N_i \geq K_i$, imputation effects for panelist i fall in two categories:

$k_i < \ell$: Since $\mathbb{P}(N_i > \ell \mid K_i = k_i) > 0$ for $\ell > k_i$, imputation increases reach and frequency.

$k_i \geq \ell$: For $k_i \geq \ell$ imputation does not affect ℓ+ reach, but only adds frequency.

The bottom-right panel of Figure 7 shows the estimated conditional distribution $\mathbb{P}\left(N = n \mid K = k; \widehat{\theta}\right)$ for $k = 0$ and $k = 2$:

$k = 0$: If the panel records zero visits, there is a $(100\% - 78.9\%) = 21.1\%$ chance the panelist actually has visited YouTube at least once.

$k = 2$: For positive observed counts there is a wide range of possibilities for the true counts – reflected in the flat conditional distribution (median equals 6).

Recall the panel data in Fig. 1a where each entry indicates whether $k_i > 0$ or not. With the conditional expectation we can draw a random sample from the conditional distribution $\mathbb{P}(N|K = k_i)$ and thus obtain a typical sample of how often panelists actually have visited the YouTube homepage. One random draw is shown in Figure 1c.

6.3 Reach estimation

Based on $\widehat{\theta}$ we can give a probabilistic estimate of ℓ+ reach of the YouTube homepage from desktop devices in Germany (see Fig. 8). The weighted average of the imputed conditional probabilities given the observed events k_i in the panel (see also Section 5) yields an imputed 1+ reach estimate of 36.4%. Comparing the curves in Fig. 8 shows that this is a large uplift from the empirical reach estimate of 19.4%.

6.4 Estimation and imputation by demographic group

The overall model gives good overall estimates, but advertisers are usually interested in specific demographic groups. Figure 1b showed that the empirical distribution of counts in the panel data varies greatly by demographic group, e.g., as expected younger people have a much higher observed count than older generations. However, just from the observations alone it is not immediately clear if this occurs because young people truly watch more YouTube, if they just have a lower missingness rate, or a combination of the two. In this section we apply our demographic-specific estimation techniques to answer this question.

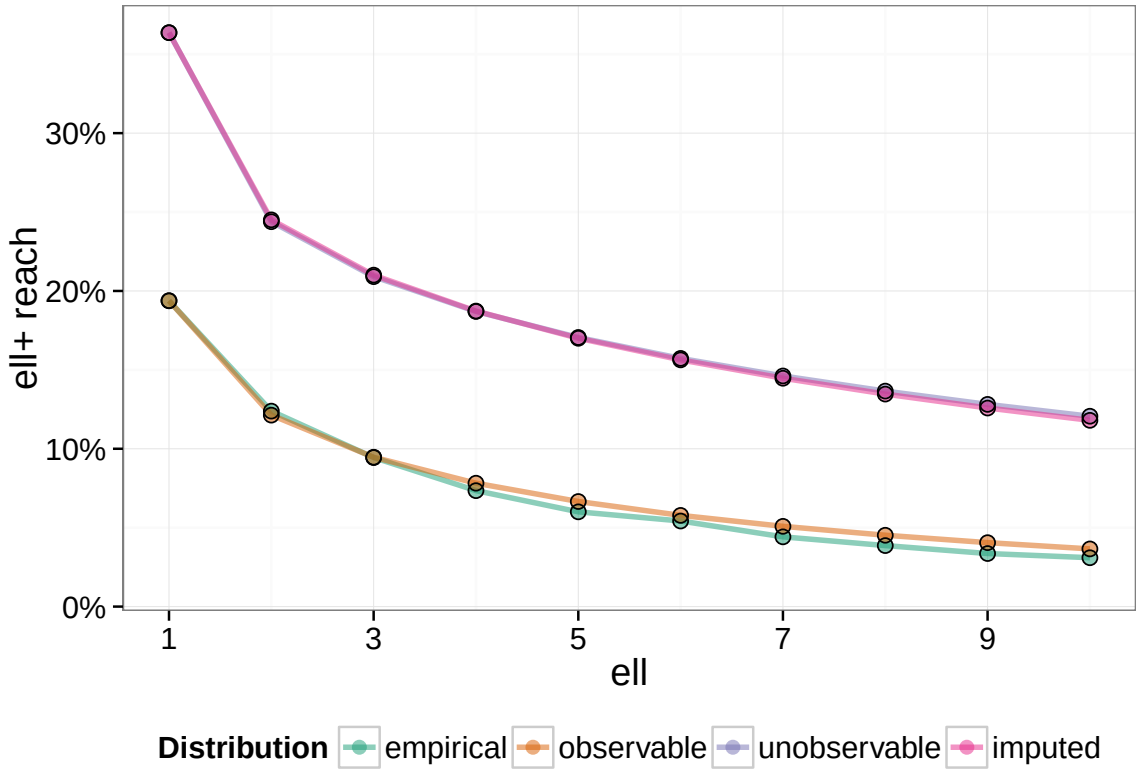


Figure 8: Comparison of $\ell+$ reach of the overall model for four types of reach estimation.

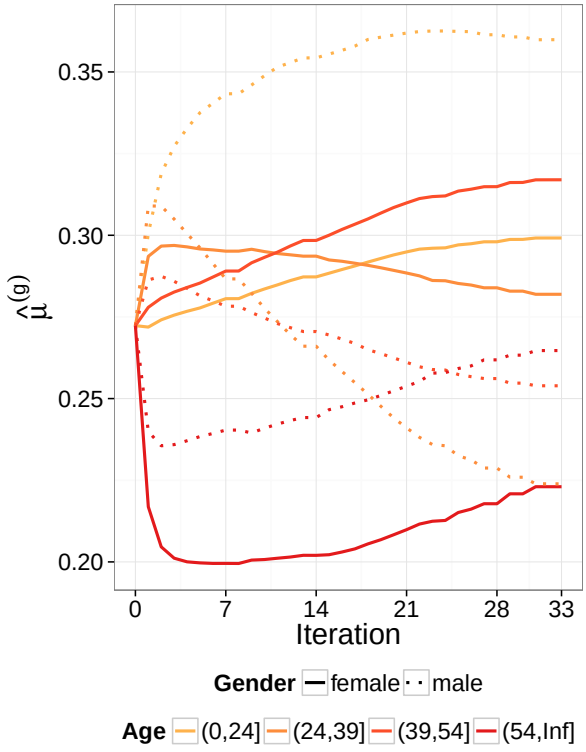
To evaluate the effects of missingness estimation by demographic group we compare the overall estimate $\widehat{\theta}$ to the model where each $\widehat{\mu}^{(g)} = \widehat{\mu}_{Logs}$ and the constrained MLE (using Eq. (18)).[9] Recall, that $\widehat{\mu}^{(1:G)}$ must average out to $\widehat{\mu}_{Logs} = 0.27$.

For the "Fix avg of mus" estimation we set δ from the smoothness penalty in (20) equal to the distance of a missingness vector with ± 0.15 around $\widehat{\mu}_{Logs}$ (such a hypothetical missingness vector yields $\delta_{target} = 0.01$); the penalty parameter $\kappa = 654$, which equals 10% of the total sample size $\sum_{i=1}^{P} w_i = P$. The iterative algorithm from Section 3.3.2 quickly converges to a heterogeneuous missingness solution, $\widehat{\mu}^{(1:G)}$ (Fig. 9a). To obtain a smooth iterative solution path we use $\lambda = 0.75$. Figure 9b shows the demographic-dependent variation of the remaining parameters. For example, $\widehat{q}_0^{(1:G)}$ tells us that the popularity of YouTube among gender and age groups.[10] The last row ($\mathbb{E}N$ from (4)) shows how many visits each demographic group truly has (on average), and clearly demonstrates that an overall model does not accurately capture how the viewing behavior depends on demographic status.

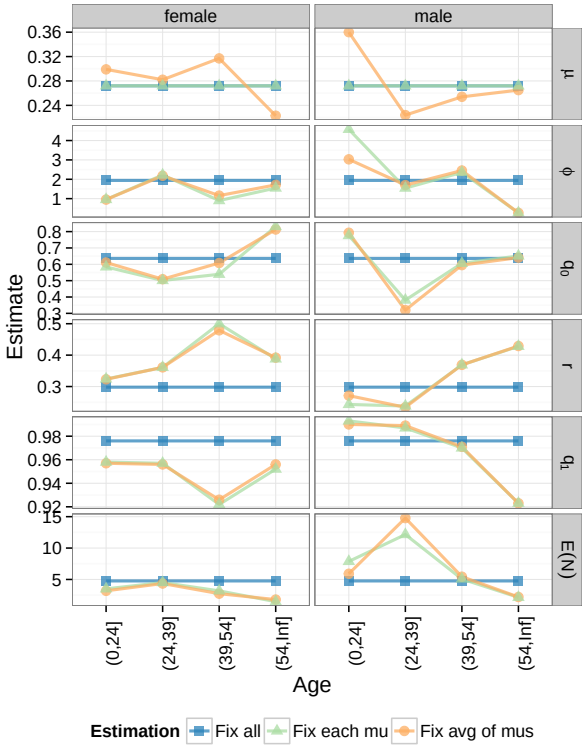
A model comparison based on log-likelihood and information criteria is given in Table 2. Group-

[9] In the figures these three models are labeled as "Fix all", "Fix each mu", and "Fix avg of mus", respectively.

[10] It seems counterintuitive that younger demographics have a higher q_0. However, recall that we analyze visits from desktop devices only. Since younger demographics heavily use mobile, a lower q_0 for older demographics is reasonable.



(a) Trace plot of $\widehat{\mu}_i^{(1:G)}$ satisfying constraint (18).



(b) Comparison of overall and demo-specific estimates. Expected number of true visits $\mathbb{E}(N)$ (last row) can be computed using (4).

Figure 9: Constrained MLE for missingness by demographic groups.

wise estimation using a fixed $\mu^{(g)} = \widehat{\mu}_{Logs}$ for each group does better than the overall model for all criteria. Letting non-missing rates vary by group increases the log-likelihood only by a small amount – which does not outweigh the additional parameters in the model according to AIC or BIC. However, as we have described above, we think that the "Fix each mu" model is too unrealistic in this web usage and YouTube visit example. We thus favor the "Fix avg of mus" solution as it has the largest negative log-likelihood including the penalty.

We evaluate the model fit by comparing the four ℓ+ reach estimates (from Section 5): i) empirical (based on $\widehat{\mathbb{P}}(K_i \geq \ell)$), ii) observable ($\mathbb{P}\left(K_i \geq \ell; \widehat{\theta}\right)$), iii) imputed ($\mathbb{P}\left(N_i \geq \ell \mid K_i = k_i; \widehat{\theta}\right)$), and iv) unobservable ($\mathbb{P}\left(N_i \geq \ell; \widehat{\theta}\right)$). While the first is a pure data-driven estimate (no model), the remaining estimates are all based on the model fit (by demographic).

Figure 10 shows that the overall model fails to provide a good fit for individual demographic groups, whereas fixing each $\widehat{\mu}^{(g)} = \widehat{\mu}_{Logs}$ as well as the constrained MLE give excellent fits. For imputation, however, the differences between "Fix each mu" vs. "Fixing avg of mus" becomes apparent in Fig. 11: demographic groups with lower non-missingness lead to higher imputed reach. For example,

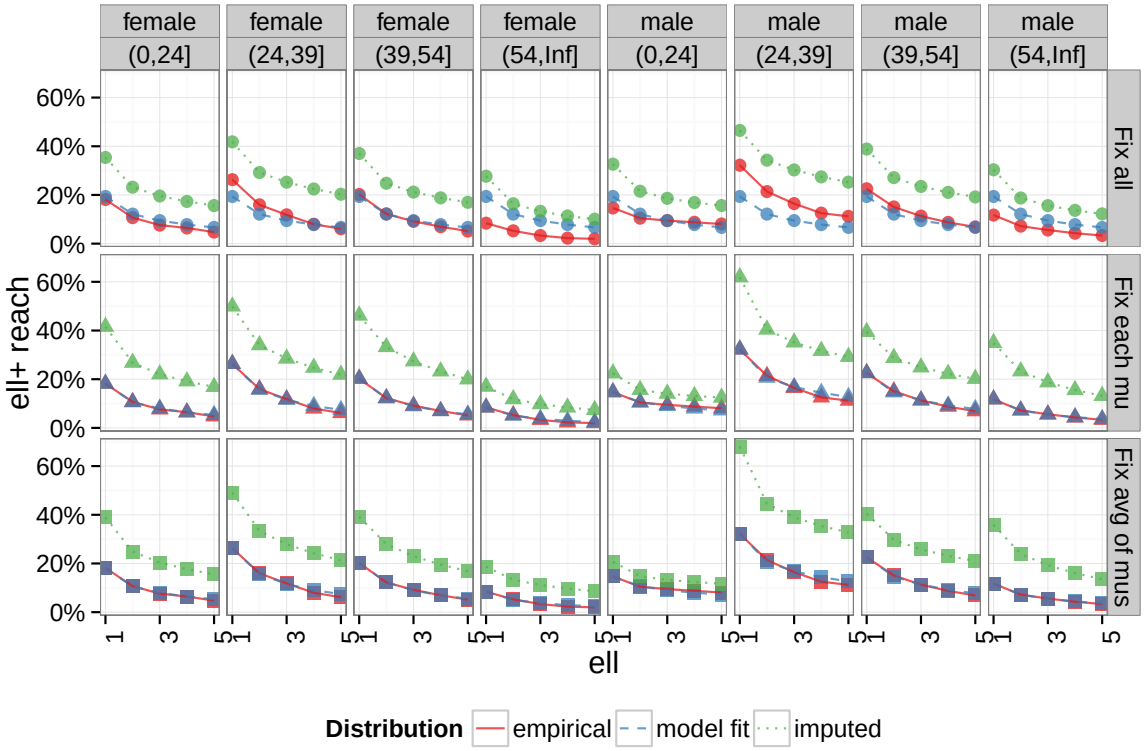


Figure 10: Estimates of ℓ+-reach by demographic groups, parameter set, and reach type.

Table 2: Model fit comparison.

	Fix all	Fix each mu	Fix avg of mus
Number of parameters	5	33	39
Neg. Log-likelihood	6,329	6,189	6,188
Penalty	4.300	4.300	0.0001
Neg. log-likelihood (+ penalty)	6,334	6,194	6,188
Sample size	6,545	6,545	6,545
AIC	12,669	12,445	12,455
BIC	12,702	12,669	12,719

$\widehat{\mu}^{(f(54,Inf])} = 0.22$ compared to the overall $\widehat{\mu}_{Logs} = 0.27$. Thus by using the "Fix avg of mus" parameters the 1+ reach estimate increases from 8.4% empirical to 18.7% imputed reach rather than just 16.9% if we had used the "Fix each mu" estimates (an additional 1.7% of imputed reach).

These percentages can be converted to absolute population estimates using the total demographic weight of the panel. The adult online and TV population was estimated at $\tilde{W} = \sum_{i=1}^{P} \tilde{w}_i =$ 50.7 million. Using the imputed 1+ reach estimates for each group (using "Fix avg of mus") we estimate that from 2013-10-01 to 2013-10-31 a total of 20.1 million (e.g., 921.2 thousand from

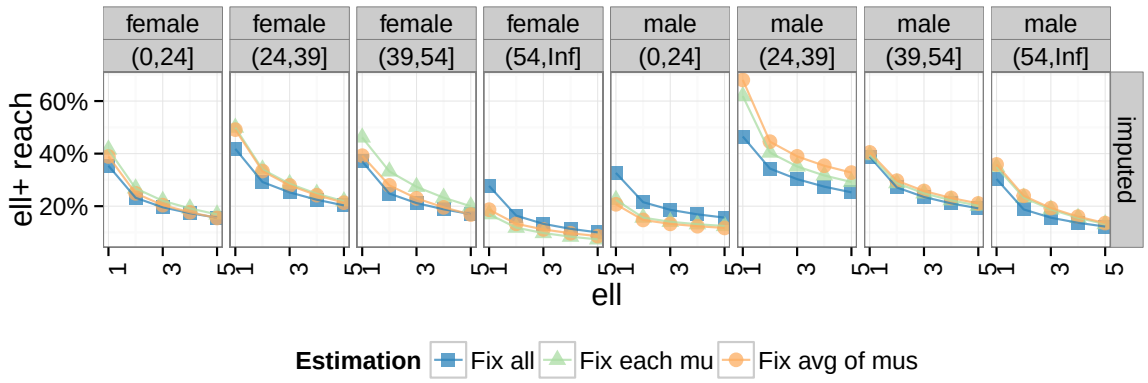


Figure 11: Imputation results by demo and for different models.

f(54,Inf]) people visited the YouTube homepage from desktop devices.

At last, we can answer the question put forward in the title of this work and estimate how many "millennials" visited the `www.youtube.de` at least once from the desktop devices. We do note that there is no consensus on the exact age range of "millennials", so we use a broad range of all adults in the (0,24] and (24,39] age groups. In this demographic, we estimate that a total of 11 million people visited the YouTube homepage portal for Germany (during October 2013).

6.4.1 Summary of demographic imputation results

Our estimates show that YouTube viewership is highly heterogeneous across demographic groups and using a "one model fits all" approach suffers from estimation bias.

The accuracy of the proposed penalized approach to estimate missingness by demographic is hard to evaluate, since the ground truth is not available. However, based on the general shape of the estimates as function of age and gender we think this approach does yield realistic estimates: first, estimates as a function of age follow a smooth pattern, and secondly, the difference between gender is merely a shift of the other curve rather than a completely different shape. Note that both the smooth shape as a function of age and the gender-shift were not forced upon by constraints in the optimization or the model, but were found automatically from the data.

7 Discussion

Motivated by the applied problem of estimating reach by demographic groups, we extend the BBNBH model to categorical covariates and propose a constrained likelihood approach to obtain unbiased imputation estimates for different demographic buckets. This method can also be used for other categorical variables (e.g., estimating weekday effects or for different economic status).

Simulations show that the BBNBH model can successfully estimate missingness and 1+ reach even when underlying truth is not available – as in most real work applications. However our simulations indicate that the method suffers from poor precision unless a good prior estimate of μ is available and $P \geq 2,000$. Further, standard errors estimated from the Hessian are smaller than expected and we suggest boot-strapping to get confidence intervals with proper coverage. We demonstrated the usefulness of our methodology to estimate how many people visit the YouTube homepage (`www.youtube.de`). For this example allowing model parameters to vary by demographic groups improved the performance.

In future work, we aim to extend the methodology to use continuous variables as predictors for $\theta^{(g)}$; in particular for $\mu^{(g)}$. Another direction for future work can focus on different penalization functions as well as a fully Bayesian approach to parametric inference. And lastly, this model treats all impressions (observed or missing) as independent while typically these impressions are tied together by cookies. Current work is focused on generalizing this modeling framework to model both cookie and impressions within cookie missingness.

Acknowledgments

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

References

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

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

[3] GfK Consumer Panels (2013). Media Efficiency Panel.

[4] Goerg, G. M., Jin, Y., Remy, N., and Koehler, J. (2015). How Many People Visit YouTube? Imputing Missing Events in Panels With Excess Zeros. Technical report, Google Inc. (`research.google.com/pubs/pub43286.html`). To appear in Proceedings of 30th International Workshop on Statistical Modelling, Linz, Austria, 2015.

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

[6] Jeuland, A. P., Bass, F. M., and Wright, G. P. (1980). A multibrand stochastic model compounding heterogeneous erland timing and multinomial choice processes. *Operations Research*, 28:255–277.

[7] Monti, G., Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. (2011). The shifted-scaled Dirichlet distribution in the simplex. In *Proceedings of the 4th International Workshop on Compositional Data Analysis*.

[8] Nielsen Solutions (2013). Nielsen Presents: Cross-Platform Home Panels.

[9] R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[10] 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.

[11] Sudman, S. (1964a). On the Accuracy of Recording of Consumer Panels: I. *Journal of Marketing Research*, 1(2):14–20.

[12] Sudman, S. (1964b). On the Accuracy of Recording of Consumer Panels: II. *Journal of Marketing Research*, 1(3):69–83.

[Wang and Koehler] Wang, X. and Koehler, J. Estimating online reach curves using logs data. In preparation for submission; to appear on `research.google.com`.

[14] Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. Econometric Analysis of Cross Section and Panel Data. MIT Press.

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

A Algorithms and Estimation Details

We use method of moments estimators to provide good, data-driven starting values for numerical optimization routines.

A.1 Iterative method of moments estimates

A.1.1 Initial estimates

Unless the non-missing rate is known, we initially set $\mu = 0.5$. For the shape of the beta prior we use $\phi = 4$ (for $\mu = 0.5$, $\phi = 2$ yields a uniform distribution om $(0, 1)$; $\phi = 4$ is slightly peaked). As for q_0, μ can be estimated from the observations and other parameters in an iterative fashion.

Estimating the hurdle parameter q_0 with the empirical frequency of zeros is biased (over-estimate), since the binomial subsampling from N can also yield $k = 0$. Below we propose an iterative procedure to reduce this bias.

For the parameters of the negative-binomial hurdle part we start with setting $r = 1$ and adjust q_1 to match (approximate) sample moments. Recall (4) and (5) which state that $\nu := \mathbb{E}N = (1 - q_0)\left(1 + r\frac{q_1}{1-q_1}\right)$ and $\mathbb{E}K = \mu \cdot \nu$. An initial estimate of q_1 can be obtained by plugging solving $\bar{\mathbf{k}} = \mu \cdot \nu(q_0, r, q_1)$ for q_1.

Based on these initial estimates we run an iterative methods of moments update to obtain better estimates for q_0 and μ.

A.1.2 Updating non-missing rate μ

The non-missing rate μ can be estimated from

$$\mu = \frac{\mathbb{E}(K \mid N > 0)}{\mathbb{E}(N \mid N > 0)} = \frac{\mathbb{E}(K \mid N > 0)}{1 + r\frac{q_1}{1-q_1}}. \tag{26}$$

The denominator can be estimated directly using initial $\widehat{r}$ and $\widehat{q}_1$ from above. The numerator, on the other hand, must be estimated using adjustments to the $K = 0$ case. First note that an estimate of $\mathbb{E}(K)$ can be obtained by $\widehat{\mu}_K = \sum_{k_i=0}^{k_{max}} f_{k_i} \cdot k_i$, where $f_{k_i} = \frac{1}{P}\sum_{i=1}^{P} \mathbb{1}(\mathbf{k} = k_i)$ is the empirical frequency of k_i. Clearly, $\widehat{\mu}_K \ll \mathbb{E}(K \mid N > 0)$ due to the zeros from a large q_0. Thus to obtain a closer estimate of $\mathbb{E}(K \mid N > 0)$, we subtract the probability of getting $k = 0$ given $N = 0$ from q_0: $\tilde{f}_{k_0} = f_{k_0} - q_0$, $\tilde{f}_{k_i} = f_{k_i}$ for $i > 0$. After re-normalizing, $\tilde{f}_{k_i} \leftarrow \tilde{f}_{k_i} / \sum_{k_i} \tilde{f}_{k_i}$, the estimate for $\mathbb{E}(K \mid N > 0)$ is $\bar{\mathbf{k}}_{N>0} = \sum_{k_i=0}^{k_{max}} \tilde{f}_{k_i} \cdot k_i$. Plugging back in (26) gives a better estimate of μ, $\widehat{\mu}_0^{(t)}$, where t is the index of iteration.

A.1.3 Updating zero-visit probability q_0

The probability of observing a zero count equals

$$\mathbb{P}(K=0) = q_0 + (1-q_0)\cdot\mathbb{P}(K=0 \mid N>0;\theta). \tag{27}$$

As $\mathbb{P}(K=0 \mid N>0;\theta)$ is independent of q_0, (27) can be re-arranged to

$$q_0 = \frac{\mathbb{P}(K=0 \mid N>0;\theta) - \mathbb{P}(K=0)}{\mathbb{P}(K=0 \mid N>0;\theta) - 1}. \tag{28}$$

An update $\widehat{q}_0^{(t)}$ can be obtained by plugging in the frequency of zeros in the observed data for $\mathbb{P}(K=0)$ and the model estimate $\mathbb{P}\left(K=0 \mid N>0;\widehat{\theta}^{(t-1)}\right)$.

As $\widehat{q}_0^{(t)}$, $\widehat{q}_1^{(t)}$, and $\widehat{\mu}^{(t)}$ all depend on each other they can be improved by iterations. By default, we use two iterations. The resulting $\widehat{\theta}_0^{(2)}$ is then used as the data-driven starting value for numerical optimization.

A.2 Bijective mapping of non-missing rates to unbounded space

We map the vector $\mu^{(1:G)}$ to the unbounded space by viewing it as a (re-scaled) vector from the probability simplex[11] $\Delta^G = \{\mathbf{p} \in \mathbb{R}^G \mid p_i \geq 0, \sum_{i=1}^G p_i = 1\}$, and using stick-breaking type transformations to obtain a bijective mapping between Δ^G and $\mathbb{R}^{G-1}$.

Formally,

$$\sum_{g=1}^{G} \frac{w^{(g)}}{W}\frac{1}{\mu^{(g)}} = \frac{1}{\widehat{\mu}_{Logs}} \Leftrightarrow \sum_{g=1}^{G} p^{(g)} = 1, \tag{29}$$

where each $p^{(g)} = \frac{w^{(g)}}{W}\frac{\widehat{\mu}_{Logs}}{\mu^{(g)}} \in [0,1]$. The vector $\mathbf{p}$ lies on the G dimensional probability simplex. It can be mapped to $\mathbb{R}^{G-1}$ via

$$\mathbf{p} \mapsto \mathbf{s} = \{s_j = \frac{p_j}{1-\sum_{i=j+1}^{G} p_i}, \quad j = 1,\ldots,G-1\}. \tag{30}$$

Every $s_j \in [0,1]$ is a cumulative fraction with respect to rest of the vector (conditional probabilities), and each s_j can be mapped to $\mathbb{R}$ using the logit transform, $\text{logit}(s) = \log(s/(1-s))$.

The inverse transformation maps any $\mathbf{y} \in \mathbb{R}^{G-1}$ to a $\mathbf{p}$ on the simplex, using the inverse logit, and the multiplying out the conditional probabilities. Finally, the group non-missing rates can be obtained by multiplying by $\widehat{\mu}_{Logs}$ and dividing each entry by weight $w^{(g)}$. While this procedure guarantees a bijective mapping between $\mathbf{y}$ and $\mathbf{p}$, not every $\mathbf{y} \in \mathbb{R}^{G-1}$ yields a valid $\mu^{(1:G)}$ (every

[11] See also Monti et al. (7).

$\mu^{(g)} \in (0, 1)$). If this occurs during the numeric optimization, we simply set the log-likelihood to $-\infty$ and let the optimizer find a better $\mathbf{y}$.