Simulation Design: **Accuracy of ADBFs**

By fanci

Reviewer: jiayupeng@, huangxichen@, shengm@

Link to this document

Objective

Evaluate the accuracy of three choices of any distribution bloom filters (ADBFs): geometric (Geo), logarithmic (Log), and exponential (Exp).

Summary

- Exp BF estimates a wide range of set sizes relative to sketch size (Simulation 2).
- Recommend choosing the decay rate $a = 10$ for Exp BF (Simulation 3).
- Provide an empirical model for the variance of relative error of the Exp BF estimate ($a = 10$), given the flipping probability p, sketch size m, number of sets k, and universe size N, under independent set generator configs (Result):

$$\sigma(\textit{relative error}) \approx 2.07 \times 9.96^{2p} \times 2.52^{pk} \times m^{-0.44} \times N^{-0.018}.$$

- Predict the affordable flipping probability, given m, k, N, and the allowed relative standard error σ (Result).
    - Example: we need $p \leq 0.143$, if we require the relative std $\sigma \leq 0.1$, using Exp BF ($a = 10$) to estimate the union reach of 20 sketches, given the universe size is 10 million, and the sketch size is 1 million.

Notations and parameters

We consider multiple publishers, and each reports one (AD)BFs, in the form of a binary vector. The BF of one publisher is the *impression* of the *set* of users reached by that publisher.

N (universe size)	The entirety of users that are possibly reached (or user IDs).
n (set size)	The number of users (or distinct user IDs) in each set (see assumption A and B below).
k (# of pubs)	The number of publishers
m (sketch size)	The length of BFs (see Assumption C)
p (blipping probability)	The flipping probability of bit flipping

$\hat{u}$ (estimated union reach)	The number/cardinality of users reached by at least one publisher.
a (decaying rate)	The decay rate parameter of Exp BF.
b (geometric probability)	The probability of Geo BF

Assumptions

Throughout, we assume that:

A. All sets are of the same size

B. Each set is independently generated/drawn from the universe.

C. All BFs (from every publisher) are of the same length (sketch size).

Exploratory Analysis

We perform three exploratory experiments/simulation studies. The first experiment is in Appendix. The second experiment suggests that Exp BF outperforms Geo/Log BFs. The third simulation study suggests that setting the decay rate parameter of Exp BF to $a = 10$ yields stable reach estimates.

Simulation 2: Set Size / Sketch Size

Fix everything else: $N = 1,000,000$, $k = 3$, $m = 10,000$, $p = 0.15$, $a = 10$, $b = 0.15$.

Vary the ratio of set size over sketch size, $n/m \in \{0.5,\ 1,\ 2,\ 4,\ 8,\ 16,\ 32,\ 64\}$

Results (Figure 2): As the ratio increases,

- Log BF overestimates the reach with increasing standard deviation of estiations.
- Exp/Geo BFs are barely affected, with the former estimate having smaller standard deviation.

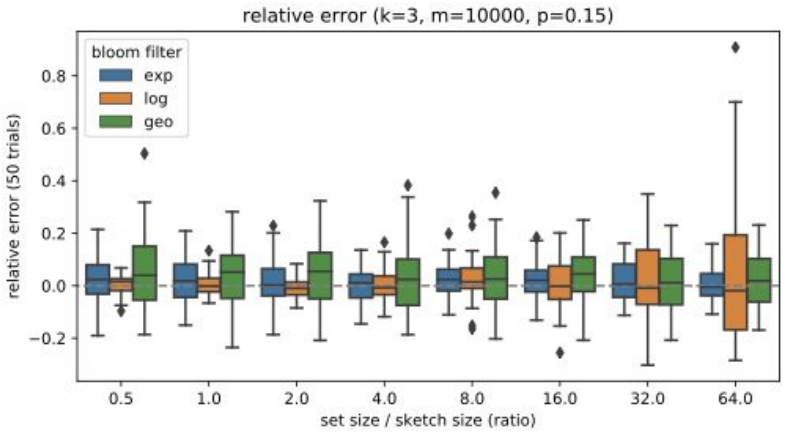


Figure 2

Simulation 3: Decaying Rate of Exp BF

This simulation concerns only the Exp BF.

Vary the decaying rate of exp BF, $a \in \{5,\ 10,\ 15,\ 20\}$ and the ratio of set size over sketch size $n/m \in \{4,\ 8,\ 16,\ 32\}$.

Fix everything else: $N = 1,000,000$, $k = 3$, $m = 10,000$, $p = 0.15$.

Results (Figure 3):

- $a \uparrow \Rightarrow$ std(relative error) is more consistent against increasing n/m.
- When n/m is small (≤ 16), $a \uparrow \Rightarrow$ std(relative error) $\uparrow$. Hence, when n/m is small, a smaller decay rate a is advantageous.
- Recommend $a = 10$.
- The same pattern holds when we change k and m (Figure 3b).
- Future research: whether the same conclusion holds for larger m and n/m.

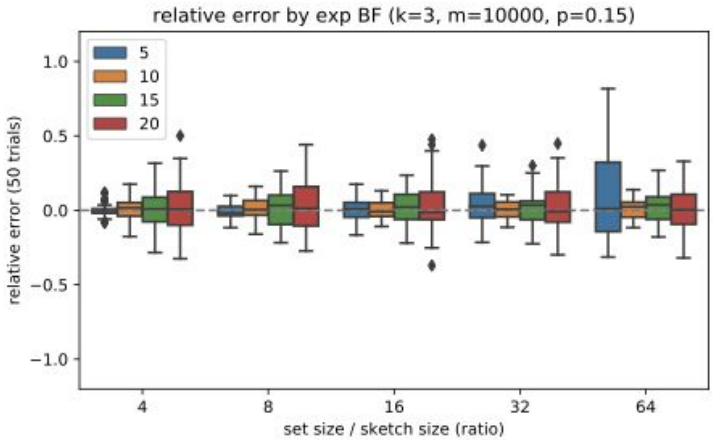


Figure 3a

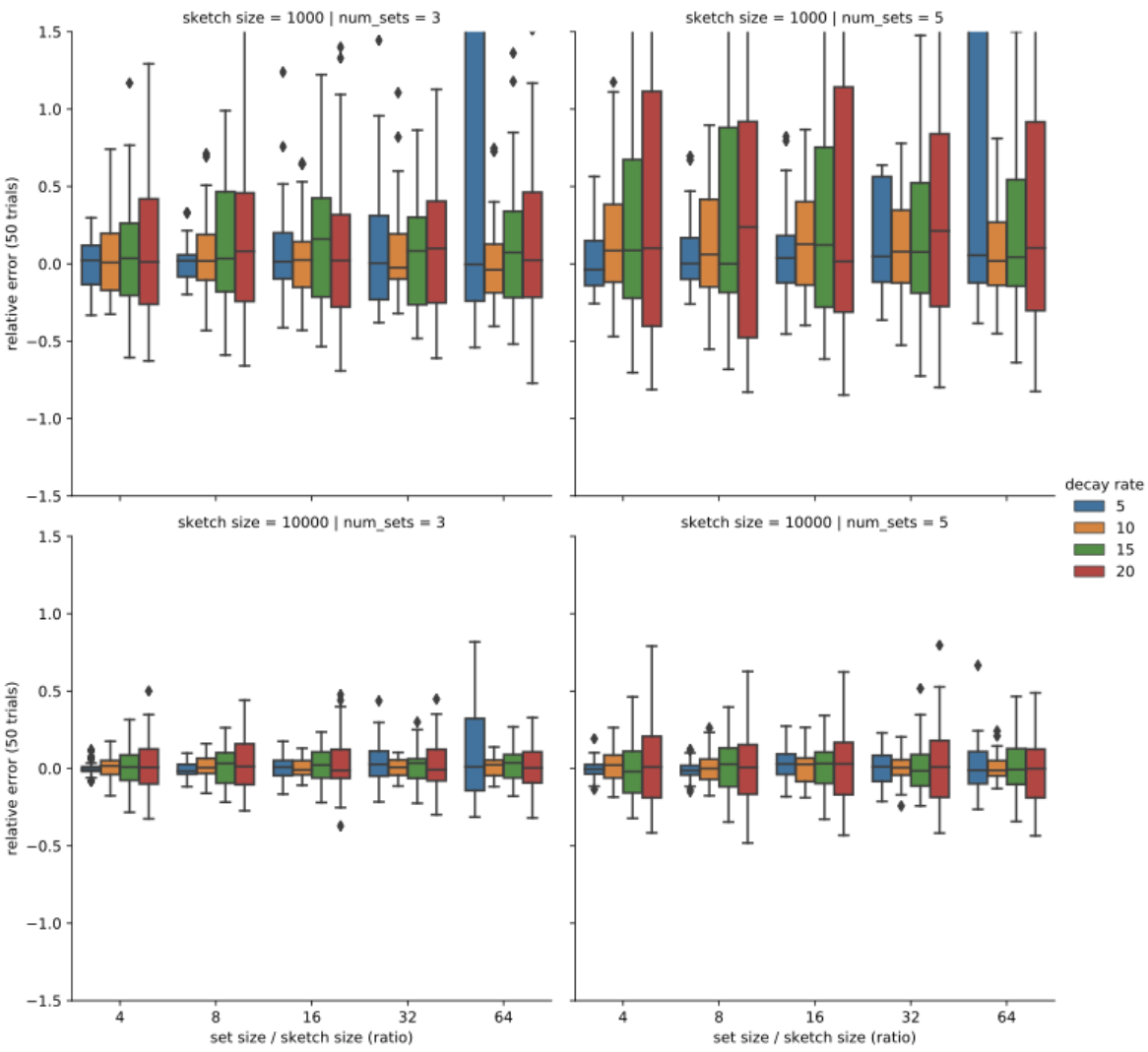


Figure 3b

Model

Let σ be the standard variance of the relative error of the Exp BF estimate for reach. We model the logarithmic transformed σ as the linear of p, $\log_{10}(m)$, k, $\log_{10}(N)$.

Experiments

- To learn such a model, we performed experiments using the following parameter grids.

Parameter	Set of choices
N	{0.1, 0.2, 0.5, 1, 2, 5} (in million)
n	20,000

k	{1, 2, 3, ..., 20}
m	{1, 2, 5, 10, 20} (in thousand)
p	{0.05, 0.1, 0.15, 0.2, 0.25}
a	10

- In above parameter grids, the set size and the decay rate of Exp BF were set to constant. This choice was suggested by Simulation 3.
- For each configuration of parameters, we repeated the experiment for 100 times, from which we calculated the (sample) standard deviation as our observations for σ. Since a small σ is of more practical interest, we also censored the data point whose observed $\sigma > 1$ in our model fitting,

  $$\sigma(\textit{relative error}) = \min\{\sigma^*(\textit{relative error}),\ 1\},$$

  where σ^* is the true value of standard variance.

- In our exploratory data analysis, we observed plausible linear relationship between $\log(\sigma)$ and the blipping probability p, as well as with sketch size $\log_{10}(m)$ and the number k of sketches (Figure 4). There is some weak evidence on the effect of N (universe size) (Figure A1).

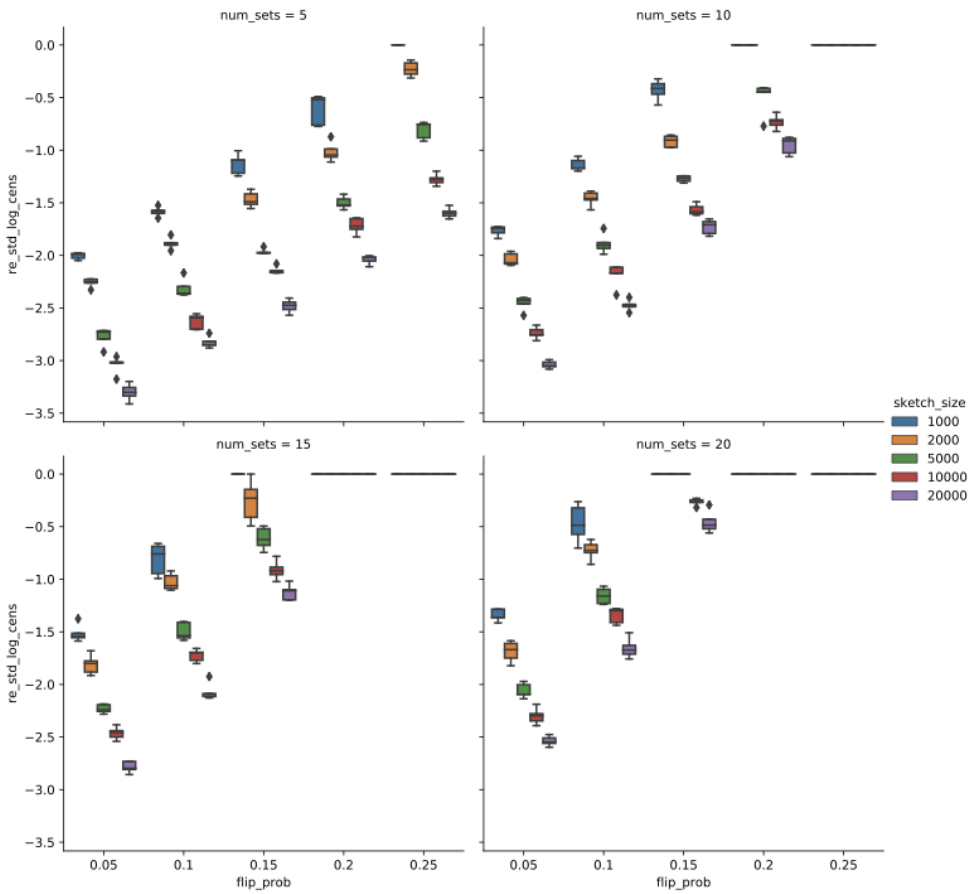


Figure 4

- Since the response variable is censored (data points with $\sigma > 1$ are coerced to $\sigma = 1$), we fitted a *Tobit* model with a censored dependent variable (with a right/upper limit)[1]. The R package `AER` (CRAN) implements a maximum likelihood based method.
    - For the Tobit model, checking the equal-variance assumption is important.
    - For comparison, the estimation using an ordinary least squares estimate is provided in Appendix.
- We also consider up to second-order interaction terms, for example the interaction between p and k. This creates a model space of 4 possible first-order terms and 6 possible second-order terms. We used a stepwise search of the optimal model that minimizes the bayes information criterion (BIC).
- The final model is

$$\log[\sigma(\textit{relative error})] = \beta_0 + \beta_p p + \beta_m \log_{10}(m) + \beta_N \log_{10}(N) + \beta_{pk}(p \cdot k) + \varepsilon$$

[1] "*The Tobit model makes the **same** assumptions about error distributions as the OLS model. In a Tobit model with heteroskedastic errors, the computer uses a bad estimate of the error distribution to determine the chance that a case would be censored.*" -- David Madigan, Professor of Statistics, Columbia University [Ref]

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ is an error term independent to other parameters. The estimated parameters and the significance are:

Parameter	β_0	β_p	β_m	β_N	β_{pk}	σ_ε
Estimate	0.726336	4.597775	-0.437599	-0.017620	0.925022	-2.039877
P-value	< 2e-16	< 2e-16	< 2e-16	1.13e-10	< 2e-16	< 2e-16

- Model diagnostics.
  - The adjusted coefficient of determination R^2 (non-censored only) is 0.9781.
  - Residuals (non-censored only) vs. fitted value (Figure 5a).
  - Q-Q plot (Figure 5b)

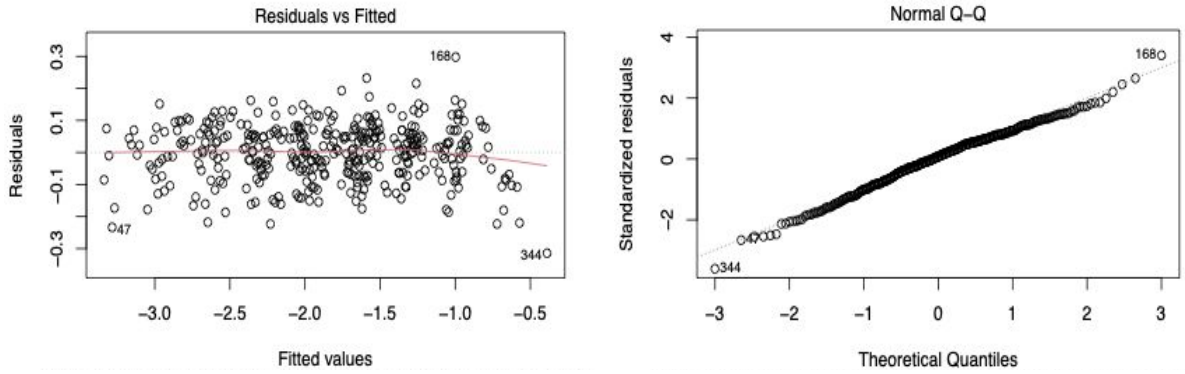


Figure 5

Interpretation and Prediction

- The model has a multiplicative form:

$$\sigma(relative\ error) = e^{\beta_0+\varepsilon} e^{p(\beta_p+\beta_{pk}k)} m^{\beta_m} N^{\beta_N}$$

Plugging in the estimated parameters, we have

$$\sigma(relative\ error) \approx 2.07 \times 9.96^{2p} \times 2.52^{pk} \times m^{-0.44} \times N^{-0.018}$$

$$\sigma(relative\ error) \approx 2.07 \times 100^{FlippingProb} \times 2.52^{FlippingProb \times NumberOfSketch} \times SketchSize^{-0.44} \times UniverseSize^{-0.018}$$

$$relative\ std \approx 2 \times 10^{2p} \times 2.5^{pk} \times m^{-0.5}$$

- Universe size: The order of N is -0.018. Although statistically significant, such small value suggests that the universe size has little effect on σ.

    - For example, increasing the universe size by 100 times shall decreases σ by about 7.8%.

- Sketch size: The order of m is roughly -0.5, which is close to the analytic results for un-noised Exp BF [Docs].

- Flipping probability and number of sketches: The effects of p and k on σ are both exponential. This suggests that in order for a reliable reach estimation (e.g., $\sigma \leq 0.1$), a sufficiently small blipping probability is critical. In the following table, we list a few predicted blipping probability using our model.

Parameter	**Scenario**				
σ (allowed)	0.1	0.05	0.1	0.1	0.1
N	10,000,000	10,000,000	100,000,000	10,000,000	10,000,000
k	20	20	20	50	20
m	1,000,000	1,000,000	1,000,000	1,000,000	500,000
$\hat{p}$ (affordable)	*0.1430664*	*0.112793*	*0.1450195*	*0.06494141*	*0.1293945*

Appendix

Simulation 1: Flipping Probability

Vary the flipping probability $p \in \{0.1,\ 0.15,\ 0.2,\ 0.25\}$.

Fix everything else: $N = 1,000,000$, $n = 5,000$, $k = 3$, $m = 10,000$, $a = 10$, $b = 0.15$.

Results (Figure 1):

- Uniform BF (red) under-estimates all the union reach.
- All three methods work at $p = 0.1$ and 0.15.
- Geo BF (green) fails at $p = 0.2$.
- All the three BFs fail at $p = 0.25$.

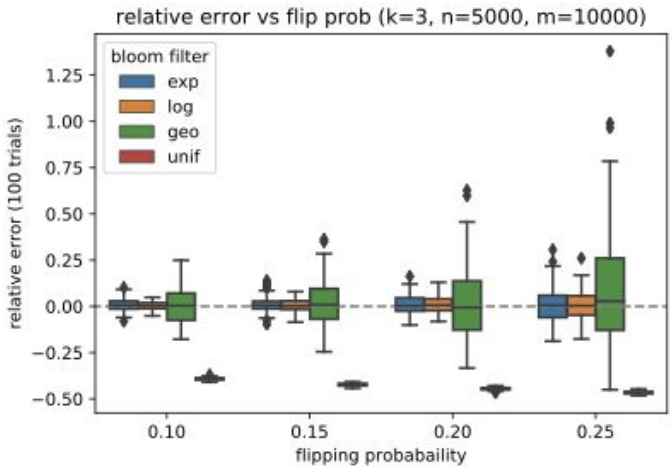


Figure 1

OLS estimation when the dependent variable Is Censored

A limitation of this approach is that when the variable is censored, OLS provides inconsistent estimates of the parameters, meaning that the coefficients from the analysis will not necessarily approach the "true" population parameters as the sample size increases. See Long (1997, chapter 7) for a more detailed discussion of problems of using OLS regression with censored data.

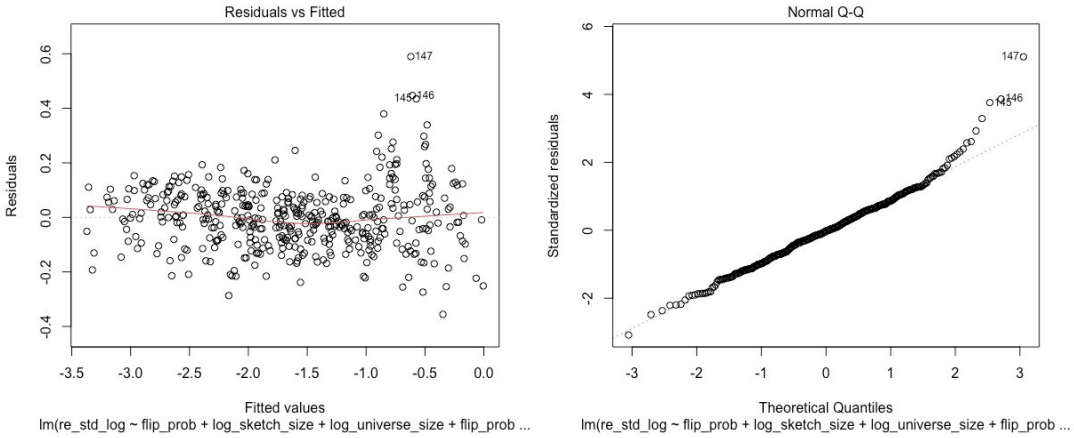
Using OLS (after discarding all censored data points), the estimated model parameters are:

Parameter	β_0	β_p	β_m	β_N	β_{pk}	σ_ε
Estimate	0.671531	4.615269	-0.419188	-0.020190	0.854098	-2.039877
P-value	1.06e-15	< 2e-16	< 2e-16	9.89e-05	< 2e-16	< 2e-16

- Multiplicative form:

$$\sigma(\textit{relative error}) \approx 1.96 \times 101.01^{p} \times 2.35^{pk} \times m^{-0.41} \times N^{-0.02}$$

- Model diagnostics:
  - Adjusted R-squared: 0.9789
  - Residual plot and Q-Q plot of residuals.



Supporting Figures

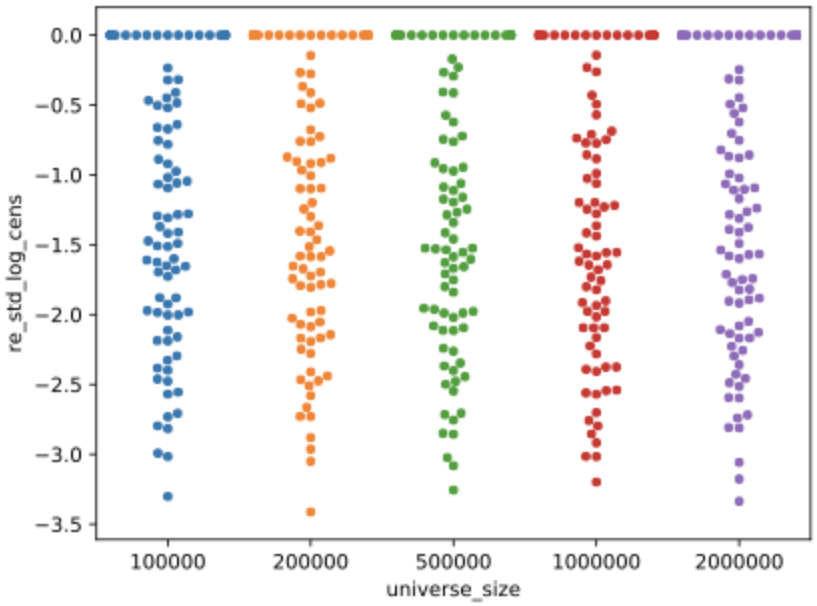


Figure A1

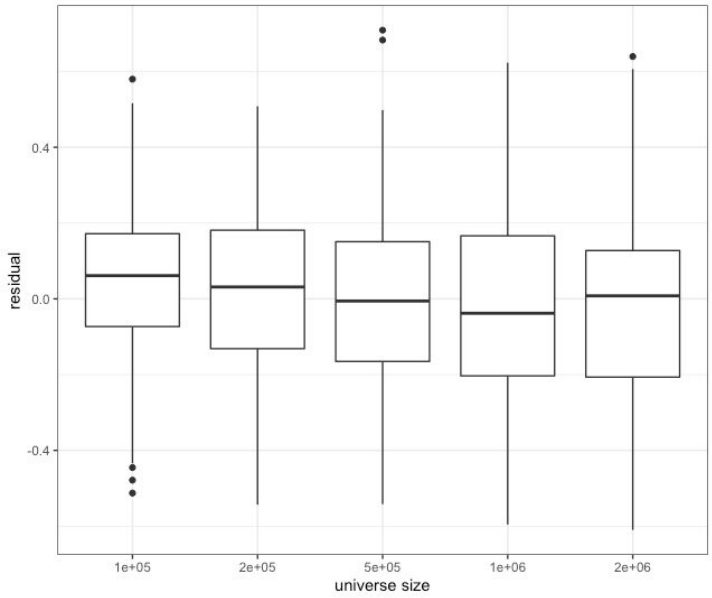


Figure A2. The residuals are estimated by fitting a model without the universe size term.