UNIVERSITY OF TARTU
Faculty of Social Sciences
School of Economics and Business Administration

**Reimo Rebane**

# Exploring a Novel Estimation Method for Voter Turnout

**Master's Thesis (20 ECTS)**

Supervisors:
Margus Niitsoo
Tarmo Jüristo
Andres Võrk

Tartu 2026

## ACKNOWLEDGEMENTS

I would like to express gratitude to Margus Niitsoo for his prior work and support on this topic. Without his support, this thesis would have not been possible. I would also like to thank Tarmo Jüristo, Andres Võrk and Jaan Erik Pihel for their feedback. And finally, I would like to thank Kaur Lumiste for reviewing this work and giving detailed comments.

## AUTHOR'S DECLARATION

Artificial intelligence (AI) tools were used in this work for the following purposes:

- Find related research articles and list their main ideas
- Summarize related concepts to improve understanding
- Improve wording of some sections
- Format tables included in this thesis
- Generate parts of the program code used in this thesis

I have written this Master's thesis independently. Any ideas or data taken from other authors or other sources have been fully referenced.

# Exploring a Novel Estimation Method for Voter Turnout

**Abstract:**

In this paper, we evaluate a novel voter turnout model that combines the methods of regression with poststratification, ecological inference and Heckman selection model. This model uses survey data and aggregated election turnout data to perform small area estimation. The aim of the model is to give insights into how individual demographic groups vote. We carry out a simulation study, which runs Monte Carlo simulations on synthetic data to test how the model handles known statistical modeling issues. The results indicate that the voter turnout model, performs better than the previous state of the art model on average. The model is also tested on the data for Estonian 2023 parliament elections, with inconclusive results.

**Keywords:** Voter turnout, simulation study, ecological inference, Heckman model

**CERCS:** S170 Political and administrative sciences, P160 Statistics, operation research, programming, actuarial mathematics

## Uudse valimisaktiivsuse hindamismeetodi uurimine

**Lühikokkuvõte:**

Selles artiklis hinnatakse valimisaktiivsuse mudelit, mis ühendab regressiooni koos poststratifikatsiooniga, ökoloogilise järeldusteooria ja Heckmani valikmudeli meetodid. Mudel kasutab küsitlusandmeid ja valimisaktiivsuse koondandmeid, et hinnata üksikute demograafiliste rühmade valimisaktiivsust. Viiakse läbi uuring, mis kasutab Monte Carlo simulatsioone koos sünteetiliste andmetega ning uuritakse, kuidas mudel tuleb toime teadaolevaite statistilise modelleerimise probleemidega. Tulemused näitavad, et valimisaktiivsuse mudel toimib paremini kui praegune tippmudel. Mudelit kasutatakse ka Eesti 2023. aasta parlamendivalimiste valimisaktiivuse hindamiseks, kus mudeli tulemuste põhjal ei olnud võimalik selgeid järeldusi teha.

**Võtmesõnad:** Valimisaktiivsus, simulatsiooniuuring, ökoloogiline järeldusteooria, Heckmani mudel

**CERCS:** S170 Poliitikateadused, administreerimine, P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

# Contents

# 1. Introduction

Understanding and predicting voting behavior, i.e. who people vote for and why, is useful in many real-world contexts. Better information could be used by policy makers for policy decisions to design more equitable voting systems, by political parties to inform their political campaigns, or by financial markets to make predictions about the future. Like with many other social phenomena, understanding behavior in elections is not simple. Opinions of people involve their core beliefs and values, what is important and salient to them at the time, how these issues are discussed in the society and more (Zaller 1992). All these aspects in turn affect their party preference, or the party or candidate who they will vote for or if they will vote at all.

Another important aspect related to election outcomes is voter turnout, or if the person goes out to vote or not. Voting patterns differ between demographic groups (H. Hartig et al. 2023; Keeter and Igielnik 2016; Leighley and Nagler 2013). For example, the highly educated tend to vote more often than the less educated and people with high income more often than people with low income. While these patterns tend to shift over time (Doherty et al. 2024; Leighley and Nagler 2013), knowing them is still useful for making decisions in the short- to medium-term. Understanding voter turnout is an interesting problem on its own and is the focus of this thesis.

There is a long history of using surveys to gather information on political behavior and various methods have been developed to derive inferences from this data. While many improvements have been made to the methods, predicting voting outcomes remains difficult and polling errors are still a problem (Keeter and Igielnik 2016). Another source of data on elections are the official election results themselves. Depending on the country, different amounts of information are available to the public on past elections. For example, in the United States[1] and the United Kingdom[2], whether people voted or not is part of the public voter registration records[3]. In these countries, common statistical tools or machine learning methods can be used to understand voter turnout. However, in most other countries, individual level data on elections is not available. Instead, what is shared is aggregate data, like per municipality turnout rates.

---

[1] https://www.eac.gov/sites/default/files/voters/Available_Voter_File_Information.pdf

[2] https://commonslibrary.parliament.uk/research-briefings/sn01020/

[3] The parts of the registration file that are published along with the process of publishing and accessing these records depends on the election system and varies significantly between countries.

The survey data and aggregated election result data can be combined to give better estimates than just using one dataset on its own. Ghitza and Gelman (2013) uses multilevel regression with poststratification (MRP) and a standard calibration approach to estimate voter turnout. While this method already gives better estimates on individual level voter turnout behavior compared to a model that just uses MRP, there is still room for improvement. An approach that separately models which people end up in the surveys based on the Heckman selection model (Heckman 1978, 1979) and uses ecological inference (King 1997) can give better results.

The goal of this thesis is to evaluate a novel voter turnout model proposed by Niitsoo (unpublished:Appendix A) using a Bayesian statistical modeling approach. We measure how the model behaves under various conditions using synthetic data and compare it to other existing models. The model uses survey data, aggregated election result data and population census data to estimate voter turnout in different population segments. Both survey data and aggregate data, come with their own set of challenges that we need to take into account when using the data for modeling. We outline common problems and some of the solutions to these problems. We also analyze the results of the model with real election data from Estonia.

The results from the synthetic data models indicate that the novel voter turnout model, in most cases, gives estimates closer to the true values of the generated population. We find cases where it clearly outperforms the other models, but also some cases, where it does not. Overall, we show that the novel voter turnout model is a noticeable step forward from the model proposed by Ghitza and Gelman (2013). The results from the Estonian turnout model are unfortunately inconclusive with no clear winner.

The rest of this thesis is structured as follows. Related concepts in existing research are summarized in section 2. Then, section 3 details the models used in this thesis. In section 4, we describe the empirical datasets. The design of the Monte Carlo experiments on synthetic data is detailed in section 5. Afterwards, section 6 shows the results of the Monte Carlo experiments and the Estonian turnout model on real-world data. Finally, the findings are discussed and summarized in section 7.

# 2. Literature Review

In the following, we provide the background information required to understand the current state of voter turnout models. Afterwards, a brief description of model evaluation methodology is given.

## 2.1 Survey-based Modeling

Voting behavior is often modeled using survey data, which is a practical approach to collect data but poses some challenges. In addition to national estimates, we are often interested in how specific subgroups behave. For example, we would like to find the voter turnout for men in a specific municipality. This is called small area estimation. A simple approach is to take all of the observations representing the subgroup of interest and to find the estimate on these observations. However, national surveys often consist of a few thousand respondents and dividing the sample this way can result in a subsample with a very small number or even no observations. The estimate cannot be calculated if there are no observations and even with a small sample, estimates can be unreliable, having wide confidence intervals. This kind of method also relies on the assumption that non-respondents in a survey are missing at random (MAR) and can be ignored. In other words, the probability of an observation being missing is constant within a subsample.

Another approach is to fit a regression model on the whole set of survey responses that predicts an outcome based on the demographic group variables. This model estimates the effect of each demographic variable and these effects can be combined to get the estimates for all possible subgroups, or cells, of the demographic variables. The regression estimates are unbiased if non-respondents are missing completely at random (MCAR) and the survey is a representative sample of the population. In practice, however, survey responses tend to be non-representative of the overall population, with some demographics over- and others under-represented. This problem of non-representativeness is further amplified by decreasing response rates to surveys (C. K. a. H. Hartig 2019; Inc 2018). If the survey data is missing at random (MAR) and if national census data is available, the estimates can be improved by adjusting them based on weights proportional to the subgroup in the census. This weighting process, after fitting a model, is called poststratification. This approach works well when there are a lot of observations. However, the estimates for the cells that contain a few to no survey observations are likely to be biased. In this case, a multilevel model can be used to get better estimates using partial pooling. Partial pooling is an approach that is a weighted average of no pooling, where each group is assumed to be statistically independent of others, and complete pooling, where we assume the groups to

have distributions that are fully specified by global features. In practice, partial pooling means that the estimates for groups with a large number of observations are close to the mean of these observations and the estimates for groups with a low number of observations are adjusted towards the overall average estimate. This combined approach of the different methods is commonly called multilevel regression with poststratification (MRP) (Park et al. 2004).

Ghitza and Gelman (2013) make improvements to the standard MRP method. To find small area estimates for voter turnout, the MRP estimates are adjusted such that the predicted outcome would exactly match the known election turnout margins of a completed election. This method leads to better estimates compared to just using MRP as it incorporates more information. However, this is an ad-hoc approach of calibrating the estimates after a model has been fitted. In this thesis, we also look at models where the aggregated turnout data is more directly integrated into the model.

## 2.2 Ecological Inference

In cases where survey data is not available or it is deemed too unreliable, inferences can still be made solely based on the official aggregate statistics reported by the government. This approach is generally known as ecological inference (EI) - the process of inferring individual-level behavior from aggregated data. The simplest example of ecological inference can be demonstrated with a $2 \times 2$ voting example, where we consider two groups, men and women, and their decision to vote, as seen in Table 1.

**Table 1**

$2 \times 2$ *Ecological Inference Problem*

| | Voting decision | | |
|---|---|---|---|
| Group | Vote | No vote | Population |
| Men | ? | ? | 1000 |
| Women | ? | ? | 1000 |
| Turnout | 1250 | 750 | 2000 |

*Note.* The inner cells are unobserved and represent individual-level behavior, while totals are observed represent the aggregate data.

What makes this problem difficult is that there are many different values for the inner cells of the table that are consistent with the aggregated totals. It may be true that 250 men voted and 750 did

not, but it could also be true that all 1000 men voted. While there was more people who voted in aggregate than those who did not vote, we cannot conclude that voter turnout among men has a similar proportion. Attributing aggregate behavior to individual-level behavior, which in this case is the group of men, is known as the ecological fallacy (Robinson 1950). It is difficult to make inferences from a single known margin. For example, the aggregated turnout numbers for a single municipality do not particularly help us understand the voting turnout within that municipality. However, knowing the per gender turnout margins for each municipality, assuming gender distributions vary between municipalities, allows us to estimate the effect of gender and municipality on turnout.

Various methods have been derived to make inferences on the unobserved cells. Earlier approaches relied on Goodman's method (Goodman 1953, 1959), which finds a regression model based on the group margins. However, because this model does not explicitly model the fact that the outcome is a proportion, it can sometimes produce estimates outside of what is possible. This problem is somewhat alleviated by only considering possibilities consistent with the aggregated numbers using the method of bounds (Duncan and Davis 1953). King (1997) greatly improved on the previous methods by using a Bayesian framework that incorporated into the optimization the fact that results need to be consistent with the totals.

The $2 \times 2$ case can be generalized to an $R \times C$ case with $R$ rows and $C$ columns. Rosen et al. (2001) show how to derive both a frequentist and a hierarchical Bayesian model for estimating voting patterns.

## 2.3 Evaluated Method

Another problem with survey data that we haven't yet discussed is, that the people who answer surveys behave differently from the people who do not. For example, people who are more interested in politics tend to answer political surveys more often and, at the same time, are also more likely to vote. In this case, the survey data is missing not at random (MNAR), which is sometimes also called non-ignorable non-response, and can lead to biased estimates. There is no general solutions to MNAR, but progress can be made by narrowing the problem with additional assumptions. Bailey (2025) use the instrumental variable approach by including a variable in the model that affects the response rate but not the outcome. Other approaches explicitly model the missingness mechanism in order to derive better estimates. One of these methods is the Heckman selection model (or type 2 tobit model) (Heckman 1978, 1979), which models the selection and outcome process separately.

We evaluate an approach proposed by Niitsoo (unpublished:Appendix A) that combines the different methods of multilevel regression with poststratification, ecological inference and the Heckman selection model, into a voter turnout model. Using the Bayesian framework, the model produces a single estimate using evidence from the different methods. The goal of this thesis is to compare this model to existing approaches in order to determine what advantages or disadvantages the model has. Due to the unpublished status of the reference paper, details of the derivation are included in Appendix A, so full context would be available to the reader. The goal is to publish both the method and it's validation as a joint paper.

## 2.4 Evaluation of Statistical Models

The behavior of the estimators of simpler models can be described analytically. However, solving complex models analytically is often infeasible. In those cases, we can instead try to understand the behavior of the model using existing real-world data. When testing statistical models using such empirical data, the true value that we want to estimate is generally unknown and often unknowable. Therefore it makes sense to turn to simulated studies using synthetic data. In this case, the true value is known, because the experimenter generated the data themselves. The design of the simulated study, of how to generate the data, what to test and what to measure has to be considered carefully.

Burton et al. (2006) laid ground work on how to plan, run and report on simulation studies. This work is later expanded on by Morris et al. (2019), who proposed the ADEMP framework. This acronym describes the steps of the framework and stands for aims, data-generating mechanisms, estimands, methods and performance measures. When describing the synthetic data experiments in section 5, we generally follow the ADEMP framework.

# 3. Turnout Models

This section provides a short comparative summary of the methodological frameworks evaluated in this study, progressing from standard estimation techniques to the proposed integrated approach.

## 3.1 Multilevel Regression and Poststratification (BP model)

Multilevel Regression and Poststratification (MRP) estimates subgroup outcomes by combining a multilevel (M) modeling stage (R) with a population aggregation stage (P).

**Binomial Regression (R).** The turnout outcome probability $\theta_c$ for a demographic cell $c$ is estimated using a probit[4] regression model:

$$\begin{aligned}
\theta_c &= \Phi^{-1}\left(X_{i,c}\beta_c^o\right) = \Phi^{-1}\left(\beta_0^o + \beta_{1[i]}^o + \beta_{2[i]}^o + \cdots + \beta_{K[i]}^o\right) \\
\beta_0^o &\sim \text{Normal}(0, \tau_0^o) \\
\beta_{k[i]}^o &\sim \text{Normal}(0, \tau_k^o),
\end{aligned}$$

where $\Phi^{-1}$ is the inverse of the cumulative distribution function (CDF) of the standard normal distribution, $\beta_c^o$ represents the turnout regression coefficients and $X_{i,c}$ denotes the demographic variables of observation $i$ in cell $c$. Because all demographic variables are categorical, $X_{i,c}$ consists of $K$ one-hot encoded demographic variables and the inner expression simplifies to a sum of the intercept and $\beta_{k[i]}$ terms, where $\beta_{k[i]}$ is the regression coefficient for a demographic variable $k \in \{1, 2, \ldots, K\}$ category $k[i] \in D_k$ of the observation $i$. The $\beta_0^o$ and $\beta_{k[i]}^o$ coefficient priors have a normal distribution with standard deviations of $\tau_0^o$ and $\tau_k^o$ respectively. In a non-multilevel model, $\tau_k^o$ values are fixed and the same for all $k$.

**Multilevel Models (M).** Hierarchical partial pooling is used to stabilize estimates in sparse cells. It is added to the model by allowing the variance of $\beta_k[i]$ to vary by demographic variable $k$:

$$\tau_k^o \sim \text{Half-Normal}(0, \sigma^o)$$

[4]Generally, the logistic link function is more common, but the difference between the models is usually not meaningful and the derivation of the more complex models is cleaner with the probit link function.

**Poststratification (P).** The cell-level estimates are aggregated to the target level $r \in R$ (e.g., a municipality) by weighting them according to the true population count $N_c$ of each cell:

$$\theta_r^{MRP} = \frac{\sum_{c \in r} N_r \theta_c}{\sum_{c \in r} N_c}$$

In our comparisons, the MRP model is called the **basic probit (BP) model** because it is the simplest model we used and all of the compared models are multilevel and use poststratification.

## 3.2 Ecological Inference (EI model)

Ecological inference (EI) reconstructs individual-level behaviors from aggregate data. In our comparisons, we use a non-standard variant of the **EI model** based loosely on Rosen et al. (2001):

$$\theta_r = \frac{\sum_{c \in r} \Phi^{-1}(X_{i,c}\beta_c^o) \cdot N_c}{\sum_{c \in r} N_c}$$
$$V_r \sim \text{Binomial}(N_r, \theta_r),$$

where $\theta_c$ and $\theta_r$ are the turnout outcome probabilities of cell $c$ and of the target level $r \in R$ respectively. Similarly, $N_c$ and $N_r$ indicate the census population counts and $V_r$ the voter counts. For more details on the derivation see subsection A.3. While not shown in the above definition, it is implemented as a multilevel model and uses poststratification.

## 3.3 Ghitza-Gelman Correction (GG model)

In their paper, Ghitza and Gelman (2013) apply a post-processing calibration step to their MRP estimates to address biases such as survey over-reporting. While not a major part of the paper, this correction makes that paper the only one we found that incorporates both the survey data as well as official aggregate statistics into a joint estimate, making this the present state-of-the-art and the approach to benchmark against. They adjust the estimated probabilities $\theta_c$ so that the aggregated total matches the official turnout count $V_r$ for a target level $r \in R$. Note that the original model used the logistic link function, whereas here we use the probit (or the standard normal CDF $\Phi$) link function.

The correction is achieved by a scalar shift $\delta_r$ that is optimized (Equation 1) and applied to all cells in the municipality (Equation 2):

$$\delta_r = \underset{\delta}{\text{argmin}} \left| V_r - \sum_{c \in r} (N_c \Phi^{-1}(\Phi(\theta_c) + \delta) \right| \tag{1}$$

$$\theta_c^* = \theta_c + \delta_r \tag{2}$$

While this approach is effective for point estimates, this method is rather ad-hoc - it is a separate step that does not propagate uncertainty or preserve the statistical properties of the original model (e.g., error bars). In our comparisons, the **GG model** is equivalent to the basic probit (BP) model that is adjusted with the GG correction. The correction is performed before the poststratification step.

## 3.4 A Novel Approach (PM and FS models)

Niitsoo (unpublished) proposed an integrated framework that combines the strengths of the models above into a unified probabilistic structure. This framework is detailed in Appendix A. Both of the models described below are multilevel and use poststratification.

**Poll and margin (PM) model.** This model combines the MRP and the EI approaches. A Bayesian model with a joint likelihood is constructed that estimates cell turnout probabilities $\theta_c$ using evidence from both approaches. Compared to the GG model, this model directly integrates the aggregated turnout information into the model.

**Full selection (FS) model.** This variant adds a Heckman selection model component to the PM model. To correct for non-ignorable survey non-response, the Heckman model constructs two latent processes for individual $i$:

$$S_i^* = X_i\beta^s + \varepsilon_i^s \quad \text{(observed as survey response } S_i \text{ if } S_i^* > 0) \tag{3}$$

$$O_i^* = X_i\beta^o + \varepsilon_i^o \quad \text{(observed as vote } O_i \text{ if } S_i = 1 \text{ and } O_i^* > 0), \tag{4}$$

where $X_i$ is a $K+1$ length (including term for the intercept) row vector for $K$ input variables of the observation $i$ and $\beta^s$, $\beta^o$ are $K+1$ length row vectors. The error terms are correlated with parameter $\rho$:

$$\begin{pmatrix} \varepsilon_i^s \\ \varepsilon_i^o \end{pmatrix} \sim \text{Normal}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

The selection process directly models which population segments are likely to respond to surveys and adjusts the outcome turnout probabilities $\theta_c$ accordingly. Additionally, because this model combines evidence from multiple approaches which provide logical bounds to the outcomes, it does not suffer from the parameter identification issues that Heckman selection models usually have. There is also an approximate version of the FS model (AFS) that uses inverse Mill's ratio (IMR) instead of sampling from a truncated normal distribution for every survey observation, which is useful if the full model fails to sample.

# 4. Data

In this section, we describe the empirical datasets used in this thesis.

## 4.1 Estonian 2021 Census Dataset

Estonian 2021 census data[5] is used as the ground truth for the demographic distribution of Estonian population. The census dataset has information on all citizens and contains categorical demographic attributes of age (7), gender (2), nationality (2), education (3) and municipality (24). For each combination of these attributes, it lists the number of people in Estonia that represented this population segment on 1st of January 2022. We filter out people who are too young to vote. We additionally filter out about 16000 people with an unknown education level (1.8%) to match the categories in the survey data. As a result, we are left with 867847 people in our reference population.

## 4.2 Estonian 2023 Parliament Election Dataset

The election turnout information is provided by the Estonian National Electoral Committee and the State Electoral Office. The results of the 2023 parliament elections are available on the official web page[6]. The general turnout for this election was 63.5%. Because the exact voter counts for age, gender and municipality categories are not available, we have to derive them using some simple calculations and approximations. To calculate the number of voters for the age and gender margins, we first use the the paper ballot proportions and paper ballot total counts per category to derive the total number of eligible voters per category. Then, using the previously calculated total eligible voters count together with the turnout percentage per category, we can find the number of people who voted for each category.

For municipalities, the available voter counts do not include people who are registered in the municipality but temporarily or permanently live abroad. Ideally, we would like to exclude voters living abroad, because the population census dataset and the survey dataset both only contain people living in Estonia. While these people make up a significant proportion (5% to

[5]For census, we use a dataset similar to RL21303 (accessed at https://andmed.stat.ee/en/stat/rahvaloendus__rel2021__haridus/RL21303). However, our dataset additionally contains population counts per municipality, whereas the former dataset only contains counts per county. This dataset was requested by SALK from Statistics Estonia.

[6] https://rk2023.valimised.ee/en/detailed-voting-result and https://opendata.valimised.ee/en

10%) of the overall eligible voters, they tend to have a low turnout rate (10.7%)[7]. However, because no information is available on which municipality these people are registered in, they are excluded from the per municipality turnout calculations. Thus they reduce the average turnout for age and gender but not municipality. Additionally, for people voting outside of their electoral district, we don't have information on which municipality these people belong to. This issue is similar to the voters living abroad, but because only a small proportion (0.5% to 3%) of voters of an electoral district fall into this category, it does not have a major effect on the results. Since every vote (including those from abroad) is cast in one regional electoral district, the votes can be localized down to that level. From there, we assume the excess votes (2.0% to 5.2% per electoral district) distribute proportionally among the counties composing the district.

To calculate the turnout for each category, we divide the voter count with the population census count for the same category. To find the margin counts, we take the per category voter counts that we already found and find the number of people who do not vote by subtracting the voter count from the population census count.

## 4.3 Estonian Survey Dataset

The survey data has been provided by SALK[8] and collected by the Norstat[9]. It consists computer-assisted telephone interviewing (CATI) samples collected in two waves, with the first from February 1st to 6th 2023 and the second from March 1st to 9th 2023. CATI samples are collected through random digit dialing (RDD), where potential respondents are called using a generated list of random phone numbers until the desired number of survey answers is reached. Non-citizens and people below the age of 18 are filtered out. We only keep the "Yes" and "No" answers to the question "Do you plan to vote/e-vote in the next elections?" (translated from Estonian *"Kas kavatsete minna valima/e-hääletada järgmistel valimistel?"*) and filter out the "No opinion" answers. Overall we filter out 248 observations and are left with 1192 responses, with 604 from the first wave and 588 from the second wave. In this dataset, 1092 out of 1192 people (92%) report that they intend to vote or already voted.

[7] https://rk2023.valimised.ee/et/participation/overall
[8] Sihtasutus Liberaalne Kodanik - https://salk.ee/
[9] https://norstat.co/

# 5. Simulation Study

This simulation study is described using the aims, data-generating mechanisms, estimands, methods and performance measures (ADEMP) framework (Morris et al. 2019).

## 5.1 Aims

The aim of this simulation study is to evaluate how the two new models (PM and FS) models defined in subsection 3.4 behave and compare them to existing models. We would like to understand when the models perform well and under which circumstances they fail. This study works with the assumption that we are working in an environment where we want to do small area estimation for a past election, so official aggregate statistics are available.

## 5.2 Data-generating Mechanisms

To limit the number of simulation parameters, we take the Estonian census dataset described in subsection 4.1 as reasonable known population. Using a real population dataset gives us realistic correlations between the different demographic variables. For the experiments that use the generated synthetic data, we simulate the survey answering and voting turnout behavior for each person in the population.

We define a data generative process, generate synthetic datasets with multiple random seeds and fit the turnout models on this data. In the base case, the data generative process is designed to create data that fits the Heckman selection model described in subsection 3.4. For each person $i$ in the census, we simulate the latent selection $S_i^*$ and outcome process $O_i^*$ variables[10]. Both of the latent processes use $K + 1$ regression coefficients for $K$ demography variables plus the intercept. The $S_i^P$ and $O_i^P$ variables are the observed binary values in the population, where the

[10]In the standard Heckman selection model, it is generally assumed that the variables used in the selection process are a subset of the variables used in the outcome process. However, in this data generative process we use exactly the same set of $K$ variables $\{X_{i,1}, \ldots, X_{i,K}\}$ for both processes.

outcome $O_i^P$ indicates if the person in the population voted (1) or not (0).

$$S_i^* = \beta^s X_i + \varepsilon_i^s = \beta_0^s + \sum_{k \in K} \beta_k^s X_{i,k} + \varepsilon_i^s$$

$$O_i^* = \beta^o X_i + \varepsilon_i^o = \beta_0^o + \sum_{k \in K} \beta_k^o X_{i,k} + \varepsilon_i^o$$

$$S_i^P = \begin{cases} 1 & \text{if } S_i^* \geq 0 \\ 0 & \text{if } S_i^* < 0 \end{cases}$$

$$O_i^P = \begin{cases} 1 & \text{if } O_i^* \geq 0 \\ 0 & \text{if } O_i^* < 0 \end{cases}$$

The error terms of the latent processes are drawn from a bivariate normal distribution with a correlation parameter $\rho_\varepsilon$. In the Heckman selection model, the variance of one of the error terms is fixed to 1 for model identification reasons. In our generative process, for simplification and without loss of generality, both of the variance terms are fixed to 1.

$$\begin{pmatrix} \varepsilon_i^s \\ \varepsilon_i^o \end{pmatrix} \sim \text{Normal} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_\varepsilon \\ \rho_\varepsilon & 1 \end{pmatrix} \right)$$

Because all of the variables $X_{i,k}$ are categorical and are one-hot encoded, the latent processes can be rewritten using an indicator function

$$\mathbb{1}_{d \in D_k}(X_{i,k}) = \begin{cases} 1 & \text{if } X_{i,k} = d \\ 0 & \text{if } X_{i,k} \neq d \end{cases}$$

$$S_i^* = \beta_0^s + \sum_{k \in K} \sum_{d \in D_k} \beta_{k,d}^s \mathbb{1}_d(X_{i,k}) + \varepsilon_i^s$$

$$O_i^* = \beta_0^o + \sum_{k \in K} \sum_{d \in D_k} \beta_{k,d}^o \mathbb{1}_d(X_{i,k}) + \varepsilon_i^o,$$

where $\beta_{k,d}^s$ and $\beta_{k,d}^o$ are regression coefficients for a demography variable $k$ category $d$ and $D_k$ is the set of categories for variable $k$. These coefficients for the selection and outcome process are drawn from a normal distribution with zero mean and standard deviations $\tau_k^s$ and $\tau_k^o$ respectively. To generate multi-level effects, the $\tau_k^s$ and $\tau_k^o$ values are drawn from a half-normal distribution with a zero mean and standard deviations $\sigma_{H,S}$ and $\sigma_{H,O}$. For identifiability reasons, coefficients for a variable $k$ (i.e. gender) over all its categories $d \in D_k$ (i.e. male, female) are restricted to

sum to zero.

$$\begin{aligned}
\tau_k^s &\sim \text{Half-Normal}(0, \sigma_{H,S}) \\
\beta_{k,d}^s &\sim \text{Normal}(0, \tau_k^s), \quad \sum_{d \in D_k} \beta_{k,d}^s = 0 \quad \forall k \in K \\
\tau_k^o &\sim \text{Half-Normal}(0, \sigma_{H,O}) \\
\beta_{k,d}^o &\sim \text{Normal}(0, \tau_k^o), \quad \sum_{d \in D_k} \beta_{k,d}^o = 0 \quad \forall k \in K
\end{aligned}$$

The coefficients $\beta_{k,d}$ indicate the true effect for each individual category $d \in D_k$ of the variable $k$. For example, a negative coefficient for age group 25-34 in the outcome process indicates that the people in that age group are less likely to vote. Similarly, a positive coefficient for women in the selection process indicates that women are more likely to answer the survey and thus be selected into the set of survey responses. Summing the effects over multiple variables gives the effect in a given population segment.

Using the generated coefficients, and the Heckman process described above, we generate the latent and observed values for each person in the population. A survey dataset of size $N_S$ is generated from the population dataset, depending on the latent selection values, by uniformly randomly sampling only from the observations where $S_i^P = 1$.

The base values of the parameters are described in Table 2. To generate data for the test cases, the base parameters are modified as necessary. In some cases, additional parameters are added with modifications to the generative process.

## 5.3 Estimands

The focus of the study is small area estimation in the context of voter turnout. Thus we are interested in estimating the population distribution of voter turnout based on the demographic variables. Using the ecological inference terminology, we want each cell of the N-dimensional tensor where each cell represents voter turnout for a specific demographic subgroup. Thus, the main metric of assessing the performance of a model is how different the estimated voter turnout margins are from the true population margins. A full turnout margin contains the turnout counts, counting those who voted and those who did not vote separately, for each cell in the population.

## 5.4 Methods

We compare the models as described in section 3, with the exact likelihood formulations provided in subsection A.2. For priors, we use semi-informative priors of $\beta_k \sim \text{Normal}(0, 3)$ and

**Table 2**

*Synthetic Data Baseline Case Parameters*

| Base Value | Brief Description |
|---|---|
| $\beta_0^s = -1, \beta_0^o = 0$ | Heckman selection and outcome process intercepts. This determines the general proportion of selection and voting turnout. For these parameter values, about 20% to 25% of the population is selected and about 50% of the population votes. |
| $\rho_\varepsilon = 0.5$ | The residual correlation between the error component of the selection and outcome process. Empirical evidence suggests that this parameter should be significant and positive (Groves et al. 2004). |
| $\sigma_{H,S} = 0.5, \sigma_{H,O} = 0.5$ | Heckman selection and outcome process coefficient variance. This determines the effect sizes in the processes. |
| $N_S = 1000$ | The number of people sampled from the selected subpopulation. The default value is a common survey size for a single-wave survey. |

$\rho_\varepsilon \sim \text{Uniform}(-0.99, 0.99)$. The analysis mainly focuses on the Ghitza and Gelman (2013) (GG) model, poll and margin (PM) model and the full selection (FS) model, but the basic probit (BP) and the ecological inference (EI) models are included where relevant.

The models are implemented in Python using the PyMC[11] probabilistic programming framework for building Bayesian models. This framework uses Markov chain Monte Carlo (MCMC) approaches to estimate the parameters of the models. The implementation of the models and the synthetic data generation code is available on Github[12].

To analyze how the models behave under different circumstances, we draw from existing literature on issues related to ecological inference methods, Heckman selection model and survey data. Using the simulation and testing process described above, we run a set of experiments. Table 3 shows the list of cases tested in this thesis. We first explore common biases, then look at known model mis-specification issues where some model assumptions may be invalid and

[11] https://www.pymc.io/

[12] https://github.com/rrebane/turnout-model-evaluation

finally test general statistical modeling problems. In order to improve readability, the exact simulation methods and their justifications of the experiments are explained together with the results in section 6.

## 5.5 Performance Measures

The main objective of the model is to recover the population margins from the data. In other words, we would like to estimate the counts of people who vote and those who do not vote for each cell of the population. Thus, we use two different metrics to calculate the distance between the margins estimated by the model and the true population margins. First, the Kullback-Leibler divergence (Kullback and Leibler 1951) is a well-known metric in statistics to measure the distance between two distributions. It is defined as

$$D_{KL}(P, Q) = \sum_{x \in \chi} P(x) \log \left( \frac{P(x)}{Q(x)} \right),$$

where $P$ and $Q$ are discrete measures and $\chi$ denotes the support of measure $P$. In our case, $\chi$ contains two counts (people who voted and people who did not vote separately) per demography cell, $P$ is the population distribution and $Q$ is the distribution estimated by the model. Note that this distance measure is not symmetrical with respect to the input distributions.

The values of Kullback-Leibler divergence are not easily interpretable. For this reason, we also use another metric called the Wasserstein distance (Kantorovich 1960; Vaserstein 1969), also known as the Kantorovich-Rubinstein metric or the earth mover's distance. Because all of the variables are categorical and we are comparing two multinomial distributions, the Wasserstein distance simplifies to:

$$D_{EM}(P, Q) = \frac{1}{2} \sum_{x \in \chi} |P(x) - Q(x)|$$

The intuitive interpretation of this measure is that it shows how much probability mass has to be moved around in the first distribution in order to make it equal to the second distribution. For both metrics, a smaller value indicates that the distributions are closer, with zero indicating equality. These distance measures are used to calculate the distance over the full turnout margins over all demographic variables.

Both of the above distance measures are calculated for the full margin and all of the one- and two-dimensional margins, where we show the average distance across all possible margins. For example, for the two-dimensional margins we show the average distance of the margins across all pairwise combinations of the demographic variables. The one- and two-dimensional margins

**Table 3**

*Model and Data Test Scenarios*

| Test Case | Brief Description |
|---|---|
| **Selection bias / non-response bias** (Groves et al. 2004; Leung and Yu 1996; Puhani 2000) | A small proportion of people agree to answer surveys, and the ones that do are non-representative of the population. |
| **Error correlation** (Puhani 2000) | High correlation between the selection and outcome process errors of the standard Heckman selection model can introduce some bias or instability. |
| **Measurement bias / over-reporting bias** (H. Hartig et al. 2023; Holbrook and Krosnick 2010) | Voting is a socially desirable behavior and people tend to over-report voting turnout behavior in surveys. |
| **Aggregation bias** (Cho 1998; Freedman 1999; King 1997) | Per group population margins correlated with the outcome margins tends to bias the estimates. |
| **Collinearity** (Puhani 2000) | Collinearity between the selection and outcome variables in the standard Heckman selection model causes identification problems. |
| **Non-normal errors** (Paarsch 1984; Puhani 2000) | The standard Heckman selection model assumes a bivariate normal error distribution. |
| **Selection and outcome effect size** | Smaller effect sizes are more difficult to estimate. |
| **Selection and outcome effect interactions** | Interaction effects increase model complexity in the number of parameters to estimate. |
| **Random noise** | More noise can produce bias and variance in the estimates. |
| **Sample size** | A smaller set of survey answers produces less precise estimates. More complex models can break down with very small sample sizes. |
| **Margin Informativeness** | A more granular aggregate margin is more informative and can be used to derive better estimates. |

are included because those are the most useful ones in terms of availability and interpretability, as turnout is usually analyzed across one dimension or two dimensional cross-tabulations. A high dimensional margin is generally no longer interpretable by humans and they are not useful in decision making.

The uncertainty bounds around the estimates are derived by running the data generative process and fitting the model multiple times while using different randomness seeds. As such, they reflect variability over different underlying datasets, not the Bayesian uncertainties of the single-dataset inference. The median of the estimates across different seeds is shown as the point estimate and the interquartile range is shown as the uncertainty interval. We chose to run each test case with 10 generated datasets in order to minimize simulation running time while still getting good estimates.

Another metric that we are interested in is the Bayesian p-value (Gelman 2003), also known as the posterior predictive p-value. This metric indicates what proportion of the estimates observed in the MCMC posterior distribution are more extreme than the reference value. It is defined as

$$p_B(y) = Pr(T(y^{\text{rep}}) > T(y)|y)$$
$$T(y) = (y - \bar{y}^{\text{rep}})^2,$$

where $y^{\text{rep}}$ represents a replicated draw from the posterior distribution, $y$ are the observed values, which in this case are the known true values from the population, and the test statistic $T$ is the square difference from the mean. When the reference value $y$ is exactly equal to the mean of the posterior draws $\bar{y}^{\text{rep}}$, the Bayesian p-value is exactly 1, as all replicated observations are more extreme than the reference value. A low Bayesian p-value essentially indicates that the true value is an outlier for the model. This metric is used to assess if the model is overconfident in its predictions. As with the distance metrics, we calculate the average metric for the turnout percentages of one- and two-dimensional combinations of the demographic variables.

# 6. Results and Analysis

In this section, we describe the results of the Monte Carlo simulations on synthetic data and the results of the empirical Estonian voter turnout model.

## 6.1 Baseline Case

We first look at the results of running the models on the synthetic data with the baseline parameters. For each of the models, we compared the estimates of the population margins of the model to the real margins of the population that we generated. The results for the dataset with the base parameters, as described in Table 2, are shown in Figure 1 and Figure 2. These two figures are the same, except for the scale used for the distance metric, with the first figure using a linear and the second using a logarithmic scale. We show this to demonstrate that using a linear scale is not practical because the difference between the largest and smallest values spans over multiple orders of magnitude. Thus, most of the results for the distance metrics in this thesis are shown on the logarithmic scale.

Figure 1. Kullback-Leibler Divergence and Wasserstein Distance With Baseline Parameters On a Linear Scale

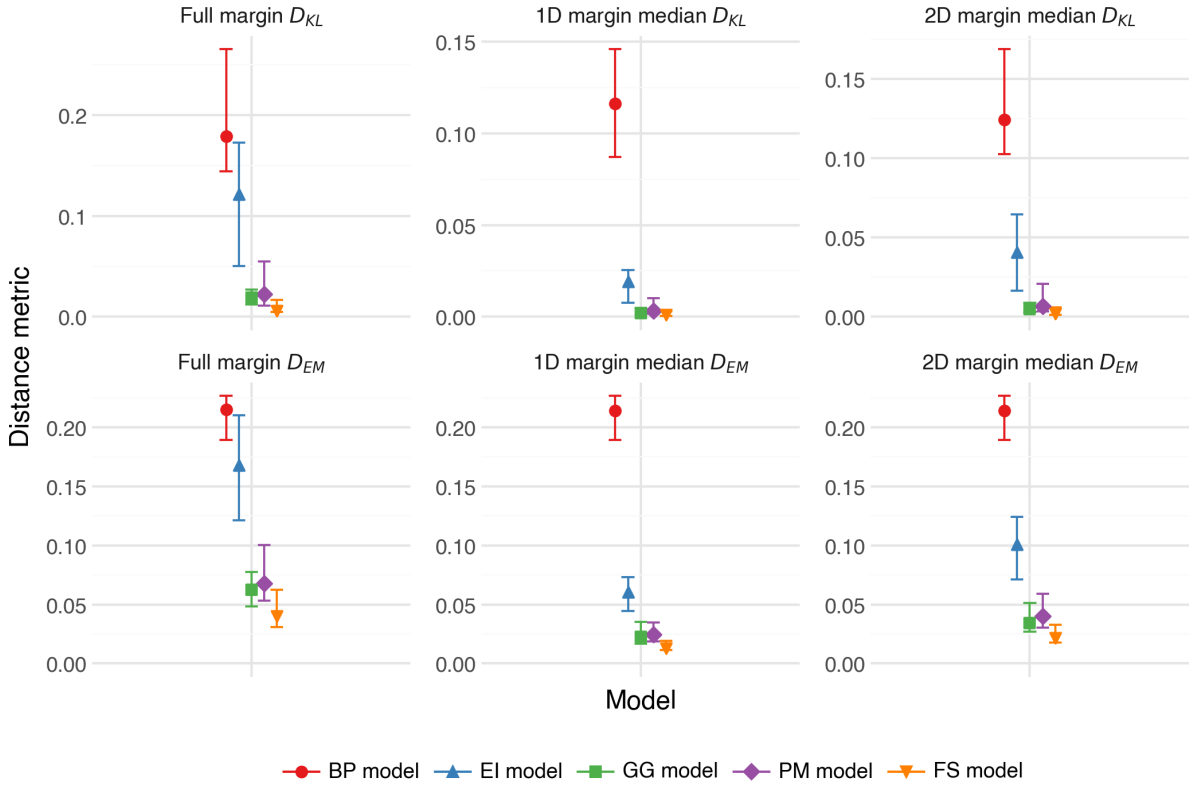


Figure 2. Kullback-Leibler Divergence and Wasserstein Distance With Baseline Parameters on a Logarithmic Scale

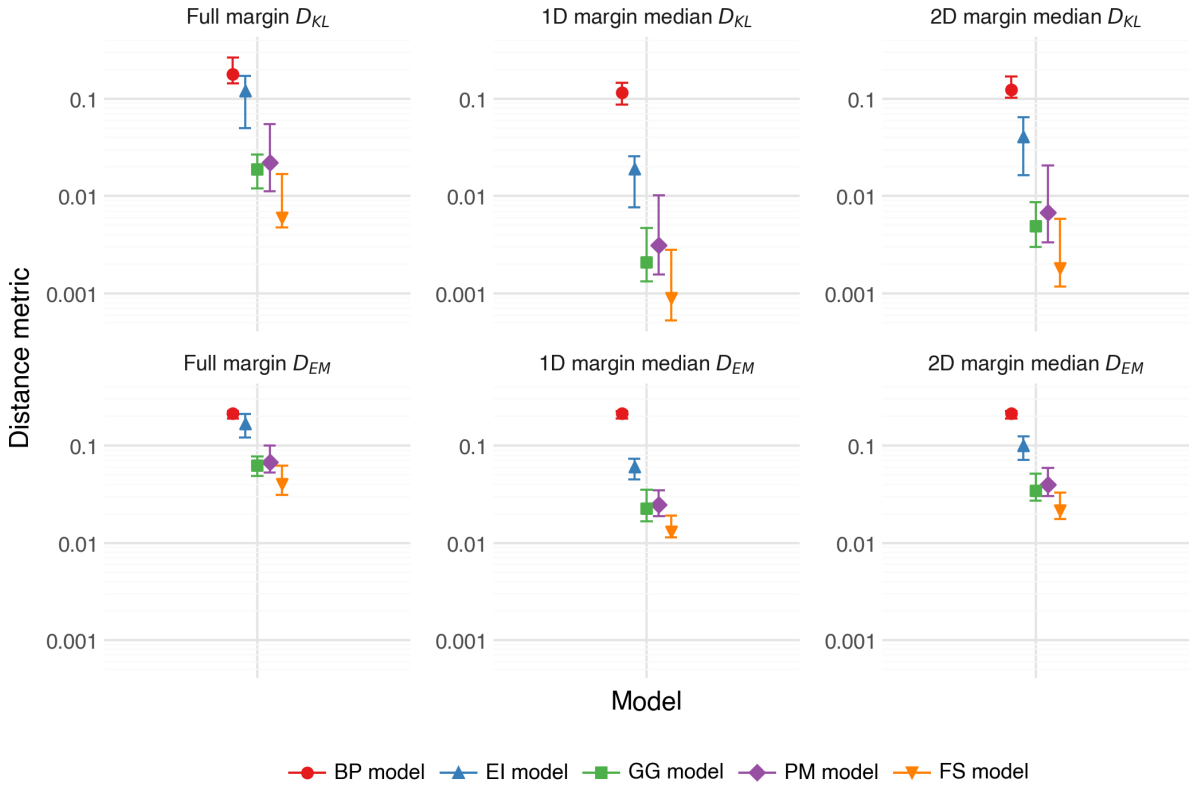
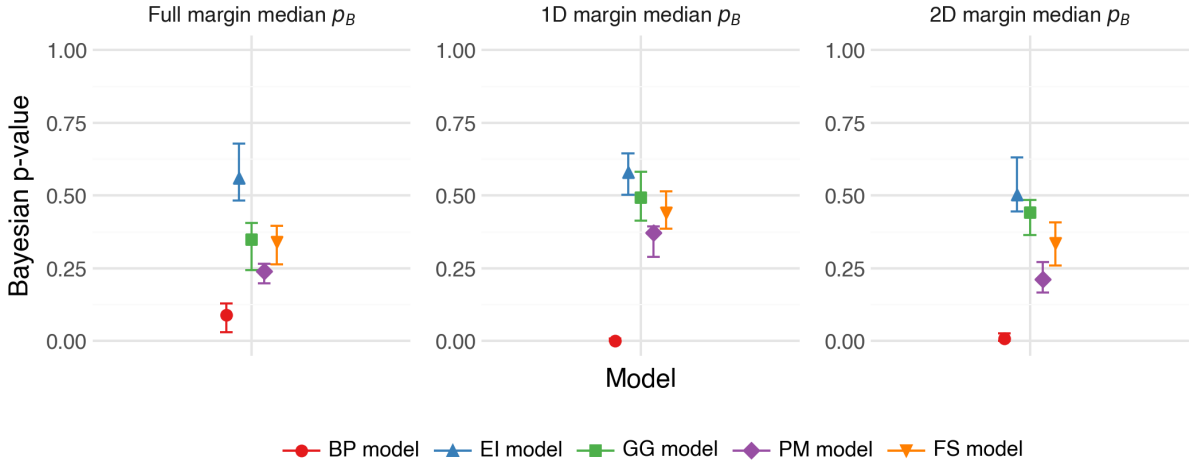



Figure 3. Bayesian p-values With Baseline Parameters



The plots display the median and the interquartile range intervals of the distance metric estimates from 10 repetitions with different seeds. A smaller distance value indicates that the model estimates are closer to the true population value. In general, the rankings of the models using the median estimates is similar between the plots. The basic probit (BP) model (defined in subsection 3.1) performs the worst, followed by the ecological inference (EI) model (subsection 3.2) which performs better than the BP model to a varying degree, depending on the inspected margin. There is a large improvement going from the basic models to the more complex models of Ghitza and Gelman (2013) (GG), poll and margin (PM) and full selection (FS) models (subsection 3.3 and subsection 3.4). This is expected, because the simpler models only work with either the survey data or the aggregated data. From the more complex models, the PM model has the largest distance, the GG model the second largest and the FS model has the smallest distance to the true population distribution. We also note that on the logarithmic scale, we cannot easily compare the confidence intervals between the models if the intervals are far apart. While the confidence interval of the BP model looks very small on the relative scale, it is the largest on the absolute scale. Generally speaking, the basic models have significantly larger confidence intervals compared to the more complex models. When comparing the complex models to each other, the confidence intervals for the Wasserstein distance are very similar between the models. For Kullback-Leibler divergence, the PM model stands out with larger confidence intervals.

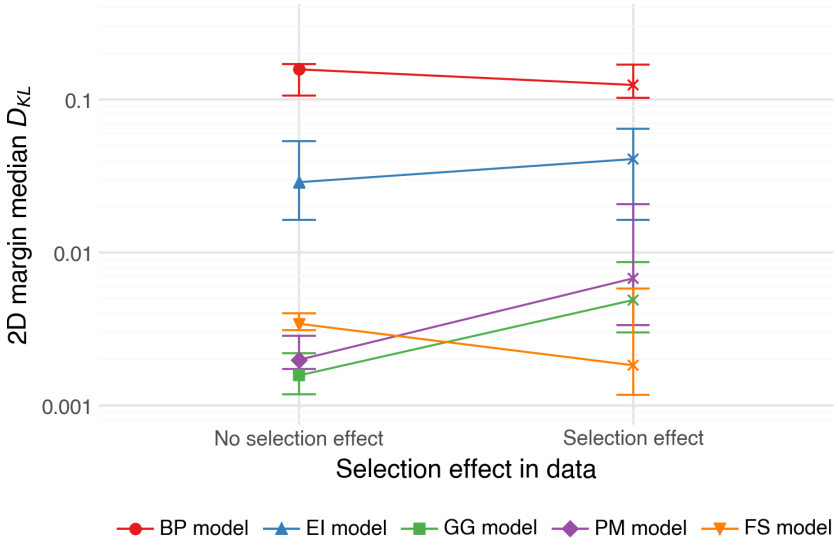
The Bayesian p-values for the one- and two-dimensional margin are shown in Figure 3. Here, a smaller value means that, according to the model, the true population margin is considered more surprising and it is less likely that the model would predict it. For all models, except the BP model, the true population margin is comfortably within a 95% high density region of the posterior. The BP and the EI model demonstrate two opposing extremes, where the former model is confident and wrong, and the latter is still somewhat wrong but not confident. The EI model gets a large Bayesian p-value because its predictions span a large set of possibilities, which also contains the true margin. The confidence intervals for the p-values of for the GG and FS models have a large overlap, whereas the p-values for the PM model tend to be a bit lower.

Going forward, as long as the different metrics agree with each other in the rankings of the models, we will only show the two-dimensional Kullback-Leibler divergence. This measure was chosen because ecological inference results are commonly validated against two-dimensional margins and Kullback-Leibler is a more common distance measure in statistics. In the following sections, we describe the results for the tested scenarios.

## 6.2 Selection Bias / Non-response Bias

Selection bias stems from the problem that the included observations are systematically different from the excluded observations. Thus conclusions drawn from the known observations do not generalize well to the whole population. For example, people who answer surveys tend to be more politically engaged and thus they are more likely to vote (Groves et al. 2004). In survey related literature, this is called non-response bias. Our data generative process produces selection bias from the Heckman selection model equation. We compare the results against the case where there is no selection bias, or equivalently, the selection process coefficients $\beta^s$ are zero and there is no correlation between the selection and outcome process or $\rho_\varepsilon = 0$.

Figure 4. Kullback-Leibler Divergence With And Without Selection Bias in Data

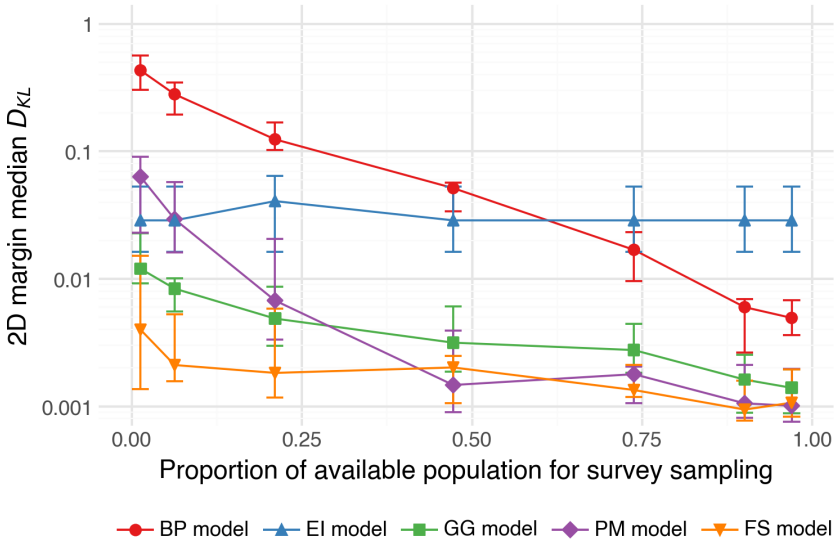



We compare the model results between the two datasets, one with and another without selection bias. The results are shown in Figure 4. Note that the baseline median distance measure using the base parameters is marked with a different the ”X” symbol for reference. The figure shows that the GG model performs better than the PM model in both datasets. Their distance measure increases in the dataset with the selection effect due to the selection bias. The FS model worse then the PM and GG model when there is no selection bias in the data and the better when there is selection bias. It is a more complex model with more parameters to estimate and likely captures random noise when there is no selection effect. The basic models of BP and EI perform worse than the other models but aren’t affected as much by the selection bias.

A high degree of censoring in the selection part of the model can increase collinearity issues and produce more biased estimates (Leung and Yu 1996). In practice, the response rates to surveys

have fallen over time and tend to be less than 10% of contacted respondents (C. K. a. H. Hartig 2019; Inc 2018). To test how the overall degree of censoring affects the results, we vary the intercept variable $\beta_0^s$ of the selection process.

Figure 5. Kullback-Leibler Divergence With Different Degrees of Censoring




As the intercept of the selection process increases, the survey responses are selected from a larger sub-population and this sub-population becomes more similar to the true population. Thus, selection bias decreases as the intercept increases. Figure 5 shows what happens when the intercept changes. The horizontal axis displays the average proportion of the population where the survey respondents are sampled from across the different data generation seeds. When the survey is sampled from a larger proportion of the population, the models give better estimates for the demography margins. Additionally, the difference between the models becomes smaller, with all of the more complex models giving similar estimates. The EI model is not affected by the degree of censoring because it does not use survey data. While a high degree of censoring negatively affects the estimates of all of the other models, the FS model handles it the best and has the lowest distance metric. We also observe that the FS model gives significantly better estimates in terms of the average Kullback-Leibler divergence measure compared to the other models already with an intercept of -3, which translates to about 1.2% of the population willing to answer the survey, and the other models catch up only around -1 (21%) or 0 (47%).

## 6.3 Error Correlation

One of the assumptions of the Heckman model is that the observations of the selection and outcome process are correlated. A positive correlation indicates that a person who answered

the survey is also more likely to vote. If the error correlation is zero, then the selection and outcome process are independent and selection bias disappears. In the simulated data, we vary the strength of this correlation $\rho_\varepsilon$ from uncorrelated to fully correlated.

Figure 6. Kullback-Leibler Divergence With Different Degrees of Correlation in Heckman Error Distribution

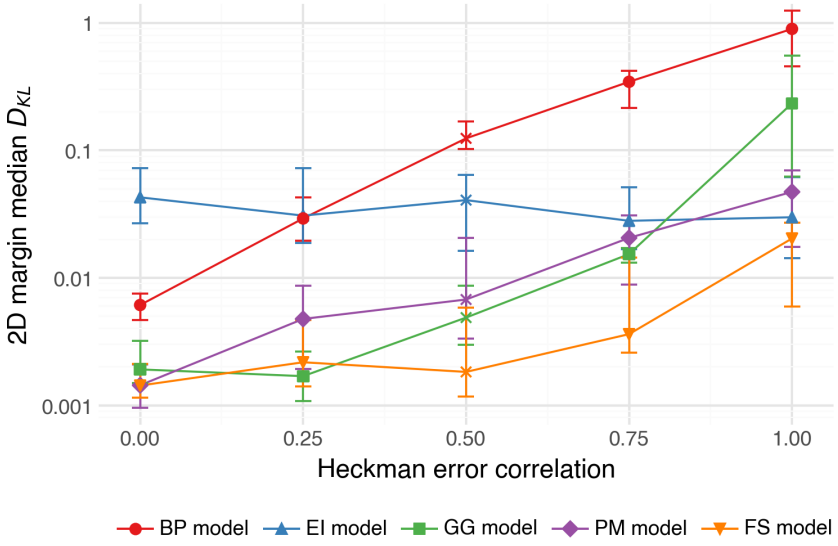



Figure 6 demonstrates that changing the error correlation affects all models similarly, except for the EI model, which is unaffected because it does not use survey data. When the Heckman error correlation becomes zero, the selection bias disappears and the models, give similarly good results. However, as the error correlation increases, the difference in how well the models are able to recover the population turnout distribution also increases. All of the models have a more difficult time to estimate the correct margins with higher correlation, especially with perfect correlation. However, the GG model has the largest slope and highest increase in the distance measure when correlation grows. The FS model performs well with moderate to strong correlation, but its advantage disappears with weak correlation values.

## 6.4 Measurement Bias / Over-reporting Bias

If the measured observation differs from what actually happened, it produces measurement bias. It has been documented that people tend to give positive answers to survey questions related to behavior that is socially desirable. Thus, questions like ”Did you vote in the past election?” and ”Do you plan to vote in the upcoming election?” tend to get a larger share of ”Yes” answers compared the actual voter turnout in the election (H. Hartig et al. 2023; Holbrook and Krosnick 2010). In survey literature, this type of measurement bias is called over-reporting bias. To

simulate this effect, we generated datasets where a different threshold is given to how people report their behavior in the survey response $O_i^S$ compared to the actual behavior in the population $O_i^P$. We do not change the actual voting turnout behavior, just the reporting in the survey, thus $O_i^P \neq O_i^S$. This is done by adding a constant bias $\beta_{OB}$ to the latent outcome variable for the survey responses, which means that some of the people close to the decision boundary who did not actually vote indicate that they did.

Figure 7. Kullback-Leibler Divergence With Different Degrees of Over-reporting

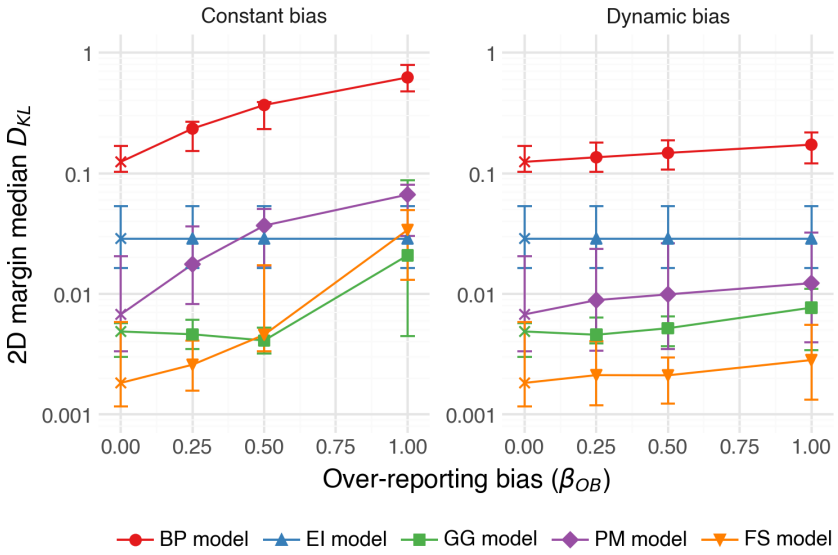



Figure 7 shows that as the over-reporting bias increases, the models have a more difficult time to recover the turnout distribution of the original population. The EI model is unaffected because over-reporting does not change the aggregate population data. With the highest tested over-reporting bias value of 1.0, where 23% of respondents mis-report their answer, the models perform no better than the EI model with distance values similar or larger. For the bias values of 0 (0%) to 0.5 (13%), the GG model does not degrade in performance, whereas the distance metric for the BP, PM and FS models increases significantly. The FS model has the smallest distance from the true population margin when no over-reporting bias is present, but it performs no better than the GG model with a bias value larger than 0.5 (13%).

## 6.5 Aggregation Bias

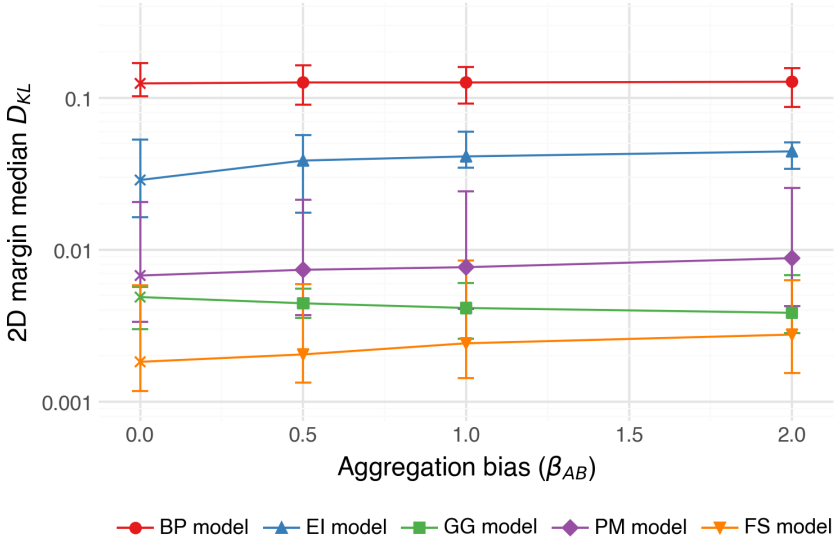
Aggregation bias is a common problem in ecological inference. It occurs when there is grouping-induced correlation between the known margins and the outcome margins (King 1997:p. 54-55). For example, sudden changes in population due to migration can raise the perceived stakes in elections and thus increase voter turnout (Chevalier et al. 2018). In this case, the per municipality

nationality margins are correlated with the turnout margins, which produces aggregation bias. Cho (1998) and Freedman (1999) have shown that aggregation bias in not uncommon in real-world data and can cause significant bias when using ecological inference methods to estimate the margins. Let $\beta_{AB,g,k,d}$ be an aggregation bias coefficient for a demography variable $k$ category $d$ grouped by variable $g$ and $w_{g,k,d}$ be the proportion of people that share the category $d$ for variable $k$, grouped by variable $g$. The latent outcome becomes:

$$O_i^* = \beta^o X_i + \beta_{AB,g,j,c} w_{g,j,c} + \varepsilon_i^o$$

To simulate aggregation bias in our dataset, we increase the per municipality $g$ turnout rates for everyone in the municipality, based on the proportion of people of the nationality "Other" (non-Estonian) category living in each municipality.

Figure 8. Kullback-Leibler Divergence With Different Degrees of Aggregation Bias



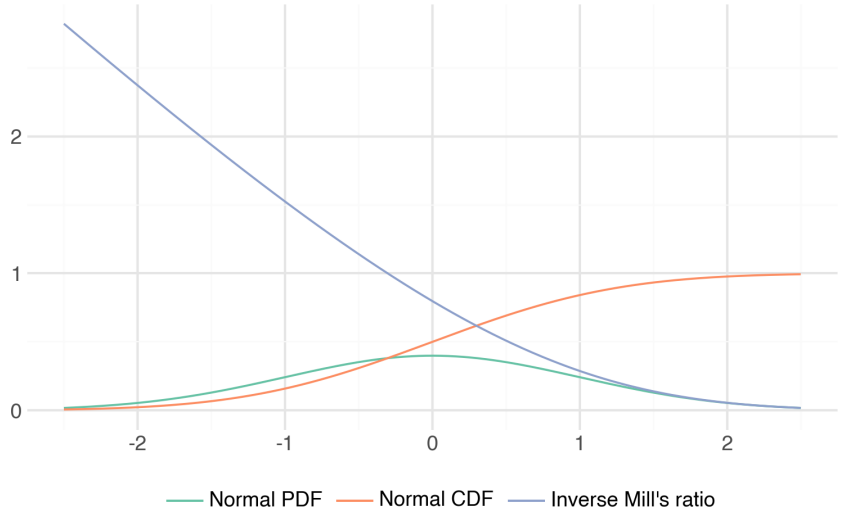

Aggregation bias can cause large biases in the estimates when using ecological inference methods. However, Figure 8 shows that in this case, all of the models handled the aggregation bias quite well. The distance metrics for all models remained at the same level or increased slightly. All in all, the performance degradation is small compared to the other test cases.

## 6.6 Collinearity

Collinearity between the selection and outcome process of the Heckman selection model creates parameter identification issues, which is a well known problem when estimating the parameters of this model. Heckman selection model parameters are often estimated using a two-step process, by first estimating the parameters of the selection process and afterwards fitting the outcome

process together with the inverse Mill's ratio (IMR) of the estimated selection process. The IMR function is the ratio between the standard normal distribution probability density function (PDF) and the standard normal distribution cumulative density function (CDF). The two-step estimation method works well as long as the IMR is non-linear. However, as shown in Figure 9, this function is near linear for a large part of its input space. If there is high collinearity between the selection and outcome process, the model cannot distinguish between the main effect and the selection effect. In this case, the outcome regression estimates become unreliable due to large uncertainty intervals.

Figure 9. Inverse Mill's Ratio Function




In real datasets, the effects of the processes can be highly correlated. For example, more educated people are more likely to answer surveys and, at the same time, they are also more likely to vote. In the standard Heckman selection model, the exclusion restriction is intended to address this problem, meaning that there must be at least one variable that is included in the selection process but not in the outcome process. Our model includes census population data and the known margins from the election, which shall eliminate this problem. To simulate collinearity, we modify the coefficients of the selection ($\beta_k^s$) and outcome ($\beta_k^o$) process by adding a term $\rho_H$ that specifies the correlation between them. The coefficients are now drawn from a bivariate normal distribution with the specified covariance matrix.

$$\begin{pmatrix} \beta_{k,d}^s \\ \beta_{k,d}^o \end{pmatrix} \sim \text{Normal}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{H,S}^2 & \rho_H \sigma_{H,S} \sigma_{H,O} \\ \rho_H \sigma_{H,S} \sigma_{H,O} & \sigma_{H,O}^2 \end{pmatrix} \right)$$

Figure 10. Kullback-Leibler Divergence With Correlated Selection and Outcome Processes

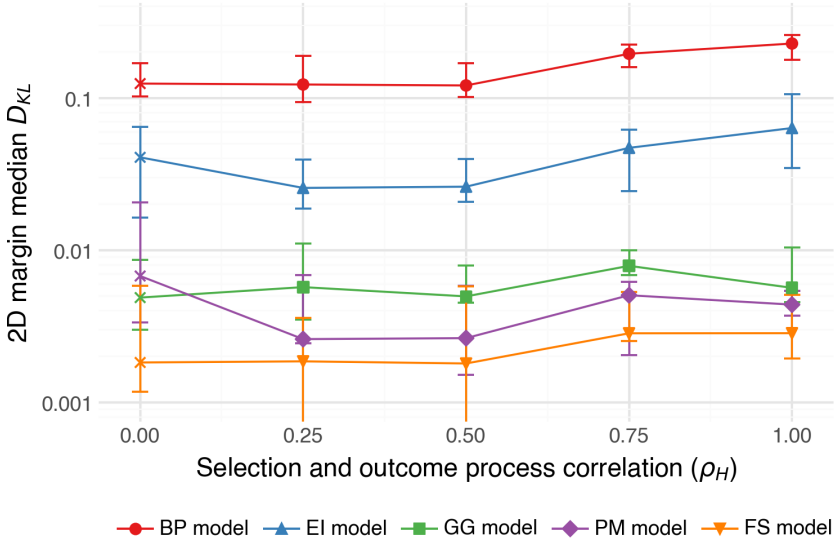


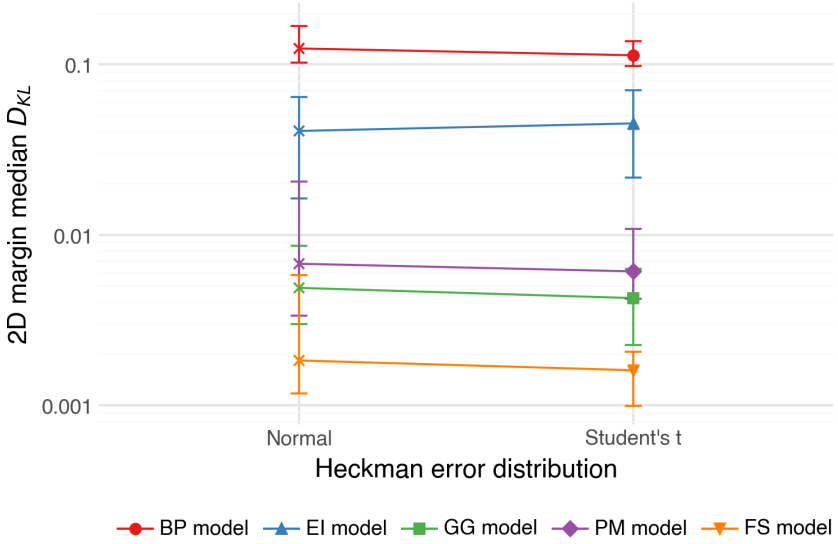

Figure 10 shows that collinearity in this case does not produce significant problems for the models. The BP and EI models show a slight upwards trend in the distance values for larger correlation coefficients. However, because the EI model only depends on the aggregate data and changing the $\rho_H$ parameter affects the survey data, then the trend for the EI model is likely due to random variance. For the other models the trend is less clear. A notable result from this test case is that the PM model handles collinearity better than the GG model and the models switch ranking if the correlation is non-zero.

## 6.7 Non-normal Errors

Heckman selection model assumes that the errors for the selection and outcome process have a joint bivariate normal distribution. Errors following a different distribution generally tend to increase the variance of the estimators (Paarsch 1984). We use a bivariate Student's t-distribution to test how the model handles an error distribution with more probability mass in the tails.

Figure 11 compares the model performance between two datasets that differ in the error distribution tails. One dataset has errors following a bivariate normal distribution and the other a bivariate Student's t distribution with 5 degrees of freedom. The plot shows that the models are robust to error distributions with more mass in the tails. The outcome variable is dichotomous and the models use a probit component to estimate the outcome, thus large values in the latent outcome process are effectively compressed and do not cause a large shift in the estimates.

Figure 11. Kullback-Leibler Divergence With Different Heckman Error Distributions

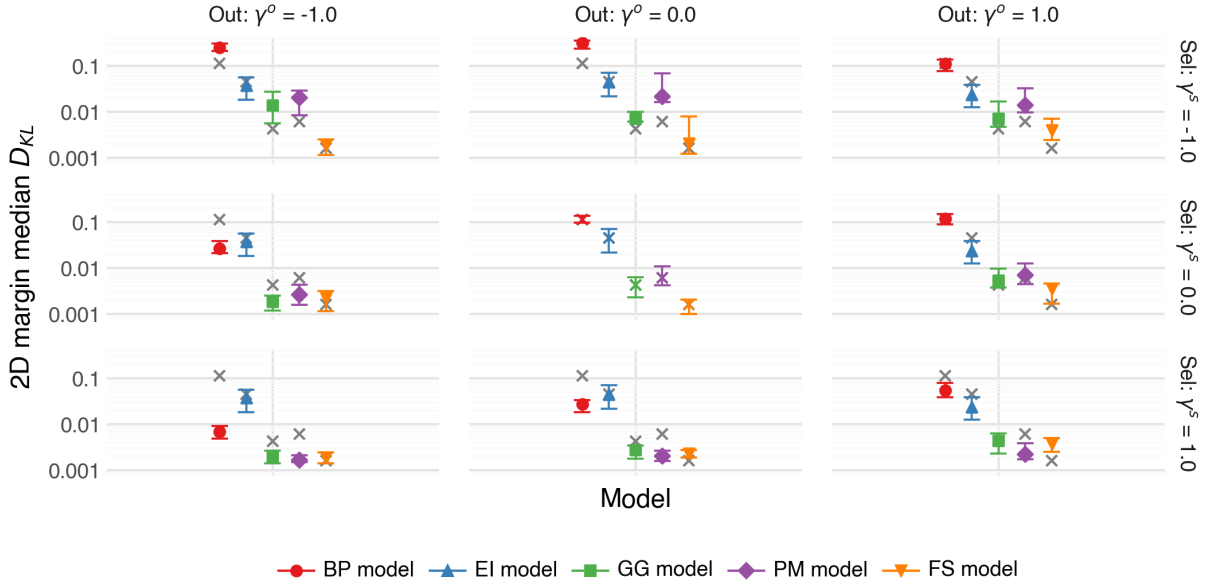



Additionally, we test how a skewed error distribution changes the results. A negative skewness parameter shifts the distribution such that the distribution has a longer tail on the negative side but the tail on the positive side falls off quickly and it works vice versa for a positive skewness. Thus, in the latent selection process, a negative skewness parameter decreases the proportion of people in the selected sub-population and a positive parameter increases it. Similarly, a negative skewness parameter decreases turnout and a positive parameter increases it. Given the degrees of freedom $\nu$, mean $\mu$, covariance matrix $\Sigma$ and skewness $\gamma$, we replace the error distribution with:

$$\begin{pmatrix} \varepsilon_i^s \\ \varepsilon_i^o \end{pmatrix} \sim \text{Skewed Student's t}(\nu, \mu, \Sigma, \gamma)$$

Figure 12 shows the results for different skewness parameters, where the Heckman selection component of the model has a bivariate skewed Student's t distribution with 5 degrees of freedom. Because selection error skewness affects selection bias, the models behave similarly to subsection 6.2. With positive skewness the distance metrics for the different models become more similar with each other, which is likely caused by the induced low selection bias. Positive skewness causes more selection bias which negatively affects the performance of all models, except the FS model. A negative outcome error skewness amplifies the effects of the selection bias and a positive skewness dampens it. The ranking of the models is relatively stable between the different cases, with the FS model performing the best, the GG model the second best and the PM the third best in terms of distance. However, for some of the cases, the model distance metrics are nearly indistinguishable.

Figure 12. Kullback-Leibler Divergence With Skewness in Heckman Error Distribution

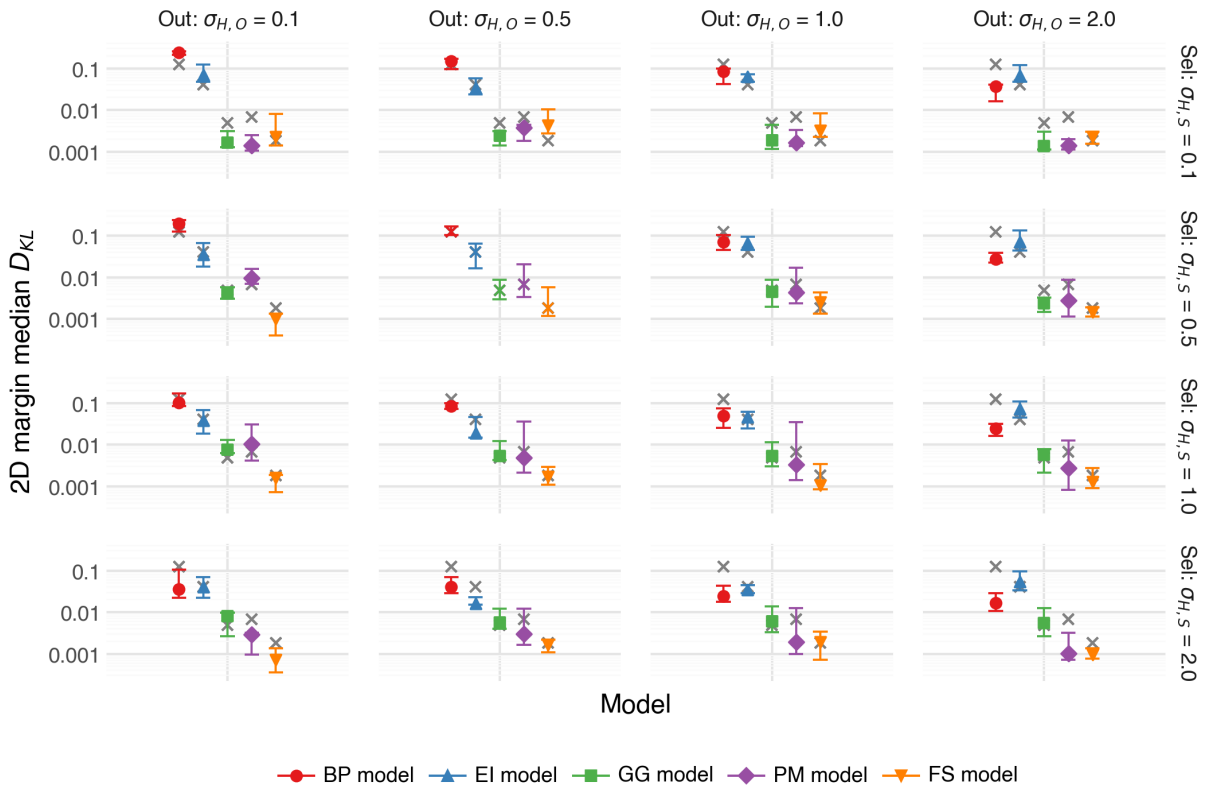



## 6.8 Selection and Outcome Effect Size

The effect sizes can be of different magnitude when comparing the selection and outcome processes. For example, people living in rural areas have the same probability as the general population to end up in the survey. However, they tend to vote at a lower rate than people living in urban areas, because in rural areas it is more difficult to access voting stations due to increased distances. In this example, the selection effect is small but the outcome effect is large and negative. In general, smaller effect sizes can be more difficult to detect. We vary the selection and outcome effect sizes independently, by modifying the parameters $\sigma_{H,S}$ and $\sigma_{H,O}$ which affect the distribution where the standard deviations of the coefficients are drawn from.

The results for different selection and outcome effect sizes is shown in Figure 13. A few general trends can be observed from the figure. All models, except for the EI model, give better estimates in terms of the Kullback-Leibler divergence measure as the outcome effect sizes increase. The FS model performs better than the GG and PM models, as long as there is non-negligible selection effect. With the smallest tested selection effect, it performs similarly or worse. In most cases, the ranking of the models is similar, with the BP model having the largest distance metrics, followed by the EI model. However, the BP model benefits from large outcome effects and switches its ranking with EI in those cases. From the other models, FS model generally performs the best and the ranking between the GG and PM model is unclear. The PM model seems to

Figure 13. Kullback-Leibler Divergence With Different Selection and Outcome Effect Sizes




benefit from larger effects. When the selection effect sizes are small, the models more complex give similarly good estimates.

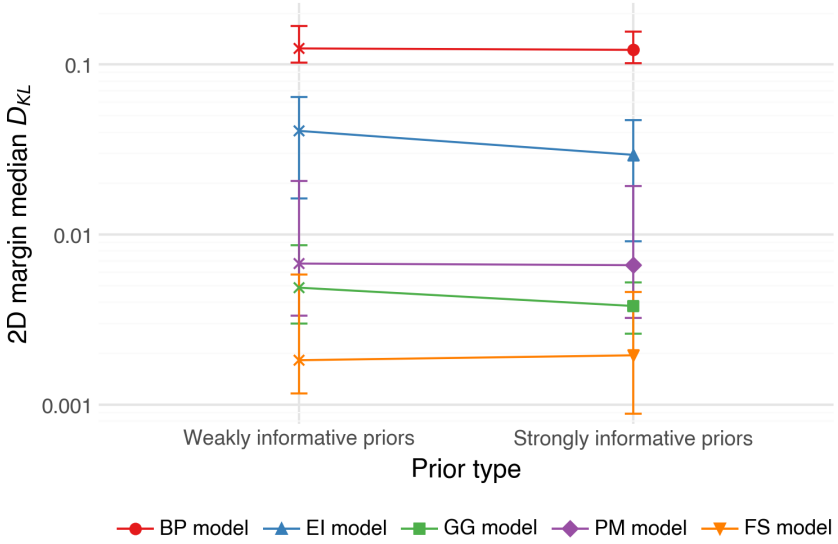
The models normally use weakly informative priors that accommodate a wide range of plausible values for the effect sizes. In practice, we sometimes have extra information that we can use to set stronger priors for the models.

Figure 14 shows the results of the models comparing weakly informative to strongly informative priors. Changing the priors has practically no effect on the BP, PM and FS models. This indicates that those models already have enough evidence to narrow down the parameter space. Using a more informative prior does seem to slightly improve the GG model and EI model distance metrics.

## 6.9 Selection and Outcome Effect Interactions

When interactions are present then an effect can depend on multiple demography variables at the same time. For example, while we expect more highly educated people to be more likely to answer surveys and to vote, it is also reasonable to expect the strength of this effect to vary over age. To add the interaction effects, we generate interaction coefficients for both the selection

Figure 14. Kullback-Leibler Divergence With Levels of Model Prior Informativeness




process and outcome process. The standard deviations of the effect sizes for the interactions are drawn the same way as the main effects. Given a set of interactions $\mathcal{I}$, the selection and outcome process become:

$$S_i^* = X_i\beta^s + \sum_{k,l\in\mathcal{I}} \beta_{k,l}^s x_k x_l + \varepsilon_i^s$$

$$O_i^* = X_i\beta^o + \sum_{k,l\in\mathcal{I}} \beta_{k,l}^o x_k x_l + \varepsilon_i^o$$

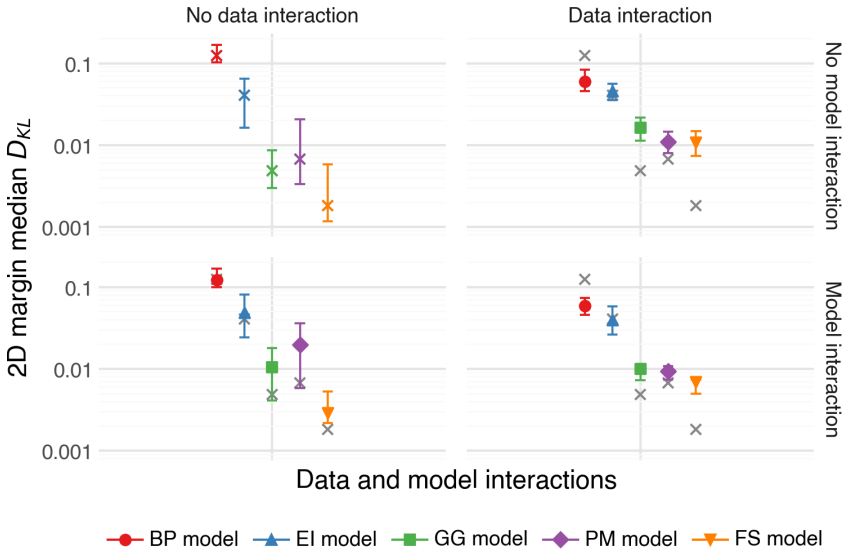
For simplicity, all pairwise interactions of the 5 input variables (age, gender, education, nationality and municipality) were added to the model.

The results for the datasets and models with and without interaction effects are shown in Figure 15. The results for the GG, PM and FS models match the intuition that a model with no interactions should be used when the data does not contain interactions and a model with interactions should be used when the data does contain interactions. The BP and EI model estimates are not affected much by the inclusion of interactions in the model. There is a penalty to the distance metric in both cases where the model does not match the data.

## 6.10 Random Noise

Real data can contain measurement error because of various reasons. For example, a phone interviewer may mishear the answer that the person provides or a person may accidentally fill out a ballot in a way that renders it invalid. We would like to test how well the model holds up when the selection and outcome process contains random noise. The noise is generated by

Figure 15. Kullback-Leibler Divergence With Interaction Effects




selecting a fixed proportion of people randomly and then determining their selection result $S_i^P$ or outcome process result $O_i^P$ by flipping a fair coin. The noise is generated independently for the selection and outcome process. Different combinations of the selection and outcome error proportions are tested.

Figure 16. Kullback-Leibler Divergence With Noise in Selection and Outcome Processes

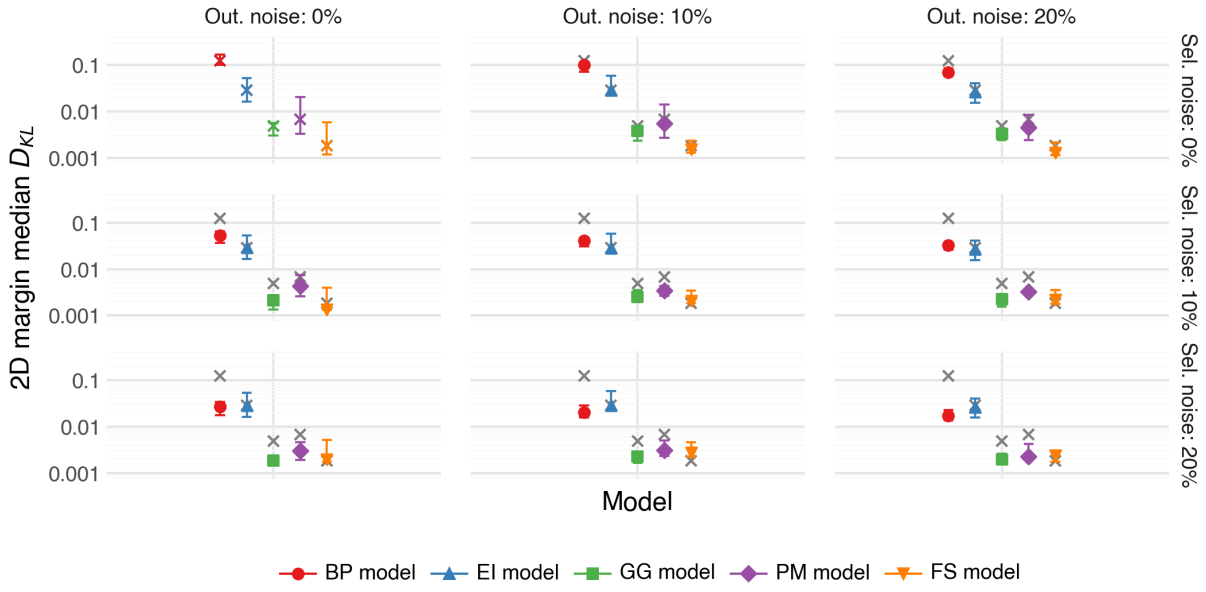


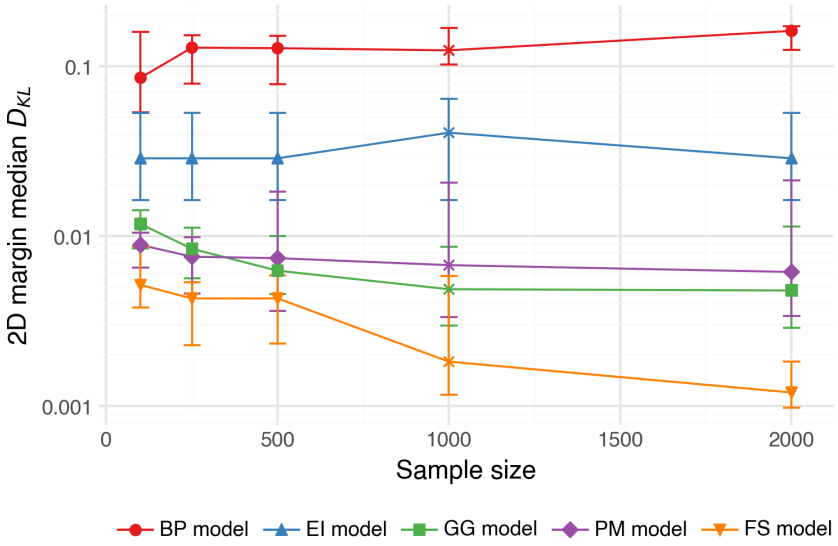

Figure 16 demonstrates that as selection noise levels increase, the distance metrics of the models become more similar to each other. This effect also exists for the outcome noise, but is much smaller. Selection noise basically sprinkles in missing completely at random (MCAR) samples,

so it is expected that selection bias becomes less relevant. As before, the EI model is unaffected because it does not rely on survey data.

## 6.11 Sample Size

As sample size decreases, we expect the models to have a more difficult time recovering the margins of the population. Additionally, we expect the more complex models to be more sensitive to small sample sizes, because they have more parameters to estimate. For this test case, we vary the sample size $N_S$.

Figure 17. Kullback-Leibler Divergence With Different Sample Sizes



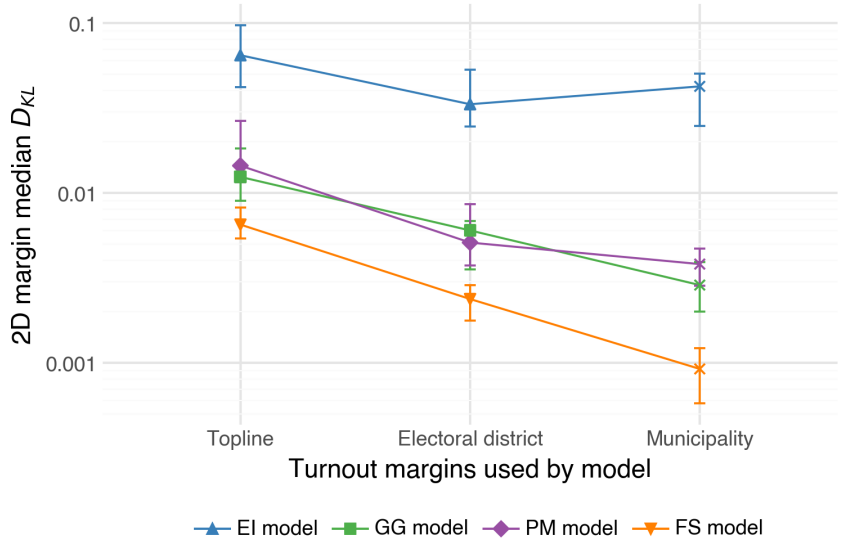

With larger sample sizes, the models are generally able to give better estimates in terms of the Kullback-Leibler divergence measure, as shown in Figure 17. The EI model does not benefit from more survey responses because it only uses aggregate data. The BP model gives slightly worse estimates with larger sample sizes, likely because it is more certain of its predictions which are based on the biased survey responses. The GG, PM and FS model all benefit from a larger sample size, but to a different degree. The decrease in the distance metric is the smallest for the PM model, then the GG model and the FS model has the largest gains. The results also show that for very small sample sizes, the PM model performs better than the GG model.

## 6.12 Margin Informativeness

Another way to improve the estimates of the model is to use more informative aggregate margins. The least informative margin is the topline which only contains the overall turnout values without any subgroups. The full margin lies at the othger extreme and is the most informative margin,

containing all subgroups of all demography variables. In practice, however, only one- or two-dimensional margins are usually available and only for a few of the demography variables. Most often, in addition to the topline margins, we know the voter turnout margins for each of the electoral districts and for each of the municipalities inside the electoral districts.

Figure 18. Kullback-Leibler Divergence With Depending on Margin Informativeness



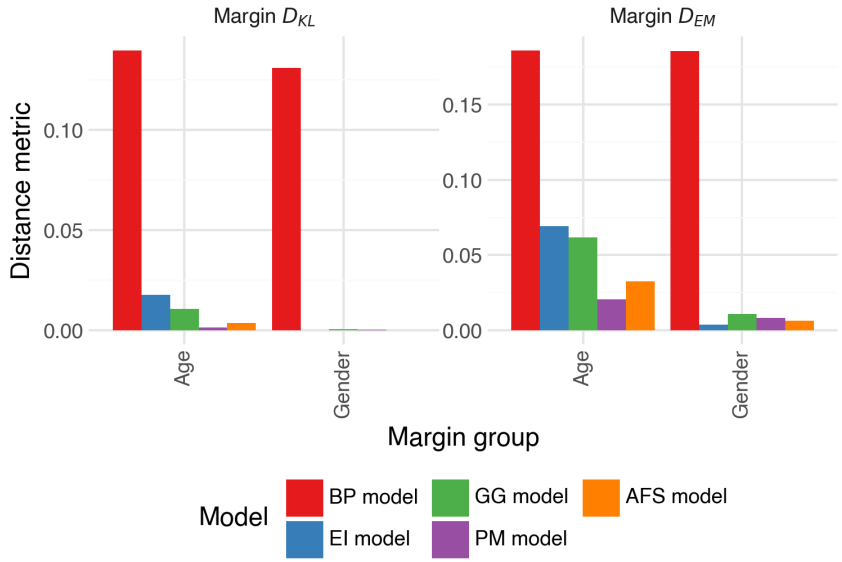

The results for the models using different margin datasets are shown in Figure 18. The GG, PM and FS model metrics improve when a more informative aggregated margins dataset is provided to the model. The EI model does not seem to benefit when using the municipality margin instead of the electoral district margin, which is unexpected and should not usually happen. The GG and PM models perform very similarly, whereas the FS model has the lowest distance to the true population margin for all tested input margins.

## 6.13 Estonian Turnout Model

To evaluate how the models behave with real-world data, we use the Estonian 2023 parliament elections as a test case. The elections took place on 5th of March 2023. The goal is to estimate the turnout margins across all of the demography dimensions using the information from the census data, the aggregated election turnout data and the survey data. On real data, FS model had a lot of divergences. While the reason for this is unclear, switching to the inverse Mills Ratio version of FS model (AFS) fixed the issue, so here we report results from that model.

The results for the Estonian turnout model are shown in Figure 19, where we calculate the Kullback-Leibler and the Wasserstein distance for the age and gender margins. Only the municipality margins were used to fit the models, the rest of the available margins serve as a

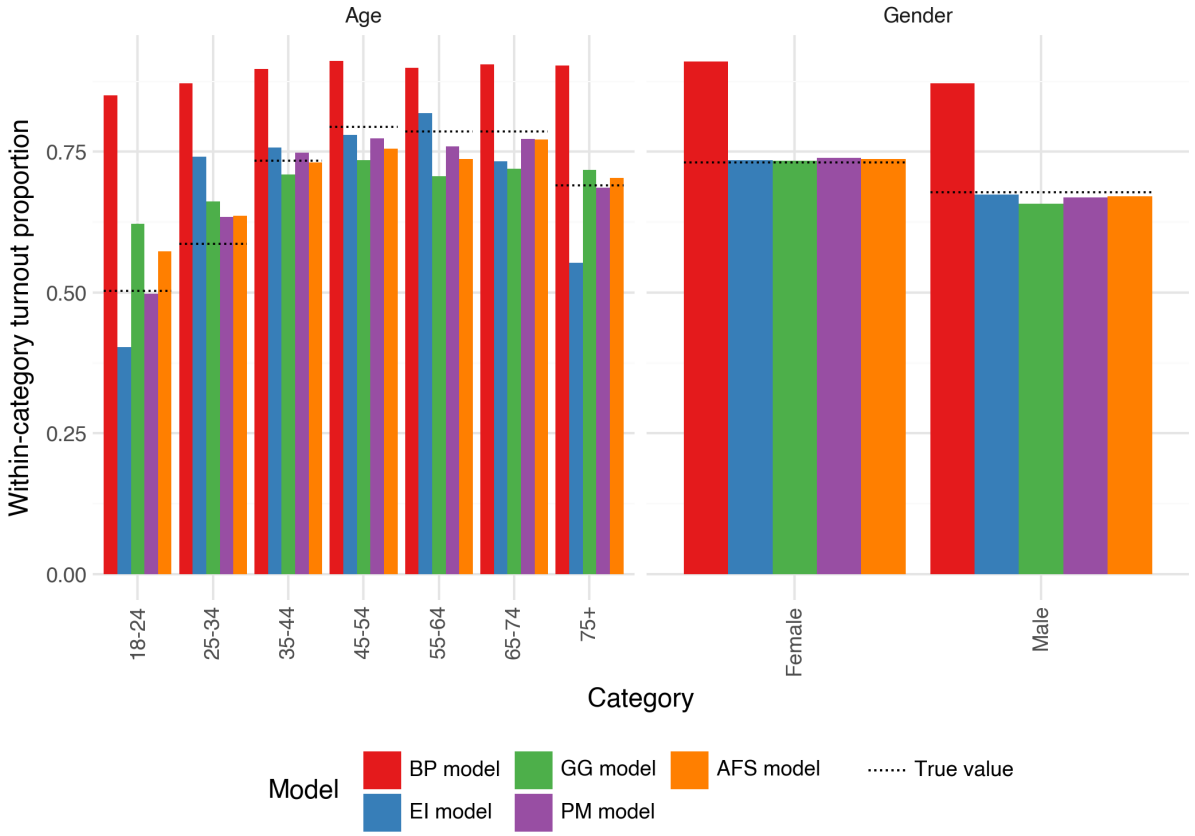
Figure 19. Estonia Turnout Model Distance Metrics




benchmark for assessing how well a model estimates the population margins. Based on the distance metrics, the margin estimated by the BP model clearly has the largest distance from the known population margin. For the age margin, the EI and GG models seem to perform worse than the PM and AFS models in terms of both distance metrics. The gender margin is estimated similarly well by all of the models.

We can also look at the individual turnout predictions of the models in Figure 20. This figure illustrates why the distance metric for the BP model is so large - all of the predictions of the model differ from the known values by a large amount. When looking at the other models, we see that the turnout predictions for the gender categories are almost spot on for all of them. For the age categories, the situation is more nuanced and none of the models consistently predicts the turnout accurately. The PM and AFS models generally give similar predictions, except for the age category 18-24, where the PM model is quite accurate and the AFS model predicts a higher turnout compared to the known value. This is also the category with the most disagreement among the models. All of the models predict a higher turnout than the true value for the age category 25-34. While these results are interesting, it is just one election with only two margins and does not allow us to make broader generalizations.

Figure 20. Estonia Turnout Model Margin Predictions



# 7. Discussion and Conclusions

We find that in the simulation study the ranking of the models is relatively stable, with the full selection (FS) model performing the best, then the benchmark GG model by Ghitza and Gelman (2013), the poll and margin (PM) model and finally, the ecological inference (EI) and the basic probit (BP) and models. The distance between the GG model and the basic models is large. The improvement from the FS model to the GG model is much smaller in absolute terms but still significant in relative terms. In the baseline case, the FS distance metrics are, on average, 13% smaller by Kullback-Leibler and 24% smaller by Wasserstein distance when compared to GG model for the two-dimensional margins.

We also saw cases where the FS model has a strong advantage over the other models. For example, in cases where there is a very large degree of censoring present in the data or if the error correlation parameter $\rho_\varepsilon$ is large. As noted in subsection 6.2, the FS model gave already good results with only 1.2% of the population accessible for polling, whereas the other models needed about 21% of the population accessible for polling to give similar results. In subsection 6.3, we found that if the error correlation between the Heckman selection and outcome process is larger than 0.5, the FS model outperforms the compared models. Furthermore, in subsection 6.11 and subsection 6.12 we showed that the additional information in the survey or the aggregate data benefits the FS model more than the other models. Overall, the FS model is therefore a noticeable step forward compared to the GG model.

This work demonstrates that the model proposed by Niitsoo (unpublished) that combines the different approaches of multilevel regression and poststratification (MRP), ecological inference (EI) and the Heckman selection model makes sense and fixes some of the problems that the individual methods commonly face. For example, the issues caused by collinearity in the Heckman selection model (Puhani 2000) are lessened because the model can draw information from the aggregate data. While there are more cases to be tested, we covered many of the known problems and outlined their effects in the voter turnout context.

The empirical Estonian voter turnout model results are unfortunately inconclusive - it covers only a single election with two known margins to compare against. Additional experiments are needed on empirical data from other elections and other countries. This remains as future work.

# References

Abrajano M., Elmendorf C. S., and Quinn K. M. (2018). The nonresponse challenge to survey research. *American Behavioral Scientist* 62.3, pp. 396–408.

Bailey M. A. (Apr. 2025). Countering Non-Ignorable Nonresponse in Survey Models with Randomized Response Instruments and Doubly Robust Estimation. en. *Political Analysis* 33.2, pp. 140–155. DOI: 10.1017/pan.2024.13. https://www.cambridge.org/core/journals/political-analysis/article/countering-nonignorable-nonresponse-in-survey-models-with-randomized-response-instruments-and-doubly-robust-estimation/AED17D9A0715AD4A15102DBD4E5B6EA8 (12/12/2025).

Burton A., Altman D. G., Royston P., and Holder R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine* 25.24, pp. 4279–4292.

Chevalier A., Elsner B., Lichter A., and Pestel N. (2018). Immigrant Voters, Taxation and the Size of the Welfare State. en. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3238550. https://www.ssrn.com/abstract=3238550 (11/11/2025).

Cho W. K. T. (Jan. 1998). Iff the Assumption Fits…: A Comment on the King Ecological Inference Solution. en. *Political Analysis* 7, pp. 143–163. DOI: 10.1093/pan/7.1.143. https://www.cambridge.org/core/journals/political-analysis/article/abs/iff-the-assumption-fits-a-comment-on-the-king-ecological-inference-solution/385588C90485A736AD1A1A7AA9614300 (11/08/2025).

Doherty C., Kiley J., and Asheer N. (Apr. 2024). Changing Partisan Coalitions in a Politically Divided Nation. en-US. Tech. rep. Pew Research Center. https://www.pewresearch.org/politics/2024/04/09/changing-partisan-coalitions-in-a-politically-divided-nation/ (12/17/2024).

Duncan O. D. and Davis B. (1953). An alternative to ecological correlation. *American sociological review* 18.6.

Freedman D. A. (1999). Ecological Inference and the Ecological Fallacy. en. *International Encyclopedia of the social & Behavioral sciences* 6.4027-4030, pp. 1–7.

Gelman A. (2003). A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing. en. *International Statistical Review* 71.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.2003.tb00203.x, pp. 369–382. DOI: 10.1111/j.1751-5823.2003.tb00203.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2003.tb00203.x (11/14/2025).

Ghitza Y. and Gelman A. (July 2013). Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups. en. *American Journal of Political Science* 57.3,

pp. 762–776. DOI: 10.1111/ajps.12004. https://onlinelibrary.wiley.com/doi/10.1111/ajps.12004 (10/23/2025).

Goodman L. A. (1953). Ecological regressions and behavior of individuals. *American sociological review* 18.6.

– (1959). Some alternatives to ecological correlation. *American Journal of Sociology* 64.6, pp. 610–625.

Groves R. M., Presser S., and Dipko S. (Mar. 2004). The Role of Topic Interest in Survey Participation Decisions. *Public Opinion Quarterly* 68.1, pp. 2–31. DOI: 10.1093/poq/nfh002. https://doi.org/10.1093/poq/nfh002 (11/11/2025).

Hartig C. K. a. H. (Feb. 2019). Response rates in telephone surveys have resumed their decline. en-US. https://www.pewresearch.org/short-reads/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/ (12/01/2025).

Hartig H., Daniller A., Keeter S., and Van Green T. (July 2023). Republican Gains in 2022 Midterms Driven Mostly by Turnout Advantage. en-US. Tech. rep. Pew Research Center. https://www.pewresearch.org/politics/2023/07/12/republican-gains-in-2022-midterms-driven-mostly-by-turnout-advantage/ (12/16/2024).

Heckman J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* 46.4, pp. 931–959. DOI: 10.2307/1909757.

– (1979). Sample selection bias as a specification error. *Econometrica* 47.1, pp. 153–161.

Holbrook A. L. and Krosnick J. A. (Jan. 2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly* 74.1, pp. 37–67. DOI: 10.1093/poq/nfp065. https://doi.org/10.1093/poq/nfp065 (11/11/2025).

Inc G. (Jan. 2018). Still Listening: The State of Telephone Surveys. en. Section: Methodology Blog. https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx (12/01/2025).

Kantorovich L. V. (1960). Mathematical methods of organizing and planning production. *Management science* 6.4, pp. 366–422.

Keeter S. and Igielnik R. (Jan. 2016). Can Likely Voter Models Be Improved? en-US. Tech. rep. Pew Research Center. https://www.pewresearch.org/methods/2016/01/07/can-likely-voter-models-be-improved/ (12/16/2024).

King G. (Apr. 1997). A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data. A Solution to the Ecological Inference Problem: Reconstructing

Individual Behavior from Aggregate Data. Princeton University Press. https://press.princeton.edu/books/paperback/9780691012407/a-solution-to-the-ecological-inference-problem.

King G., Rosen O., and Tanner M. (1999). Binomial-Beta Hierarchical Models for Ecological Inference. eng. *Sociological Methods and Research* 28.1, 61–90.

Kullback S. and Leibler R. A. (Mar. 1951). On Information and Sufficiency. *The Annals of Mathematical Statistics* 22.1, pp. 79–86. DOI: 10.1214/aoms/1177729694. https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Sufficiency/10.1214/aoms/1177729694.full (11/14/2025).

Leighley J. E. and Nagler J. (Nov. 2013). Who Votes Now? Demographics, Issues, Inequality, and Turnout in the United States. en. Princeton University Press. https://press.princeton.edu/books/paperback/9780691159355/who-votes-now (12/17/2024).

Leung S. F. and Yu S. (May 1996). On the choice between sample selection and two-part models. *Journal of Econometrics* 72.1, pp. 197–229. DOI: 10.1016/0304-4076(94)01720-4. https://www.sciencedirect.com/science/article/pii/0304407694017204 (11/11/2025).

Morris T. P., White I. R., and Crowther M. J. (Jan. 2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 38.11, pp. 2074–2102. DOI: 10.1002/sim.8086. http://dx.doi.org/10.1002/sim.8086.

Niitsoo M. (unpublished). A Novel Voter Turnout Model. Unpublished.

Paarsch H. J. (1984). A Monte Carlo comparison of estimators for censored regression models. en. *Journal of Econometrics* 24.1-2, pp. 197–213. https://ideas.repec.org//a/eee/econom/v24y1984i1-2p197-213.html (11/11/2025).

Park D. K., Gelman A., and Bafumi J. (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. en. *Political Analysis* 12.4, pp. 375–385. DOI: 10.1093/pan/mph024. https://www.cambridge.org/core/journals/political-analysis/article/abs/bayesian-multilevel-estimation-with-poststratification-statelevel-estimates-from-national-polls/22A5EF78D027E76C782B3280D400FCC9 (12/05/2025).

Puhani P. (2000). The Heckman Correction for Sample Selection and Its Critique. en. *Journal of Economic Surveys* 14.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-6419.00104, pp. 53–68. DOI: 10.1111/1467-6419.00104. https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-6419.00104 (11/02/2025).

Robinson W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* 15.3, pp. 351–357.

Rosen O., Jiang W., King G., and Tanner M. A. (2001). Bayesian and Frequentist Inference for Ecological Inference: The R×C Case. en. *Statistica Neerlandica* 55.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9574.00162, pp. 134–156. DOI: 10.1111/1467-9574.00162. https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9574.00162 (10/23/2025).

Sturgis P. and Jennings W. (2017). Trends in public opinion of electoral integrity. *Electoral Studies* 48, pp. 1–13.

Vaserstein L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problems Inform. Transmission* 5.3, pp. 64–72.

Zaller J. R. (1992). The Nature and Origins of Mass Opinion. Cambridge Studies in Public Opinion and Political Psychology. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511818691. https://www.cambridge.org/core/books/nature-and-origins-of-mass-opinion/70B1485D3A9CFF55ADCCDD42FC7E926A (01/15/2026).

# Appendices

# A. Proposed Novel Methodology

This section is unpublished work written by the supervisor of this thesis (Margus Niitsoo) and has been included here for reference.

Here we describe the model combining MRP, ecological inference and Heckman-style selection

## A.1 Data Sources

We leverage three complementary data sources to analyze voter turnout:

- **Census Data**: Complete enumeration of eligible voters stratified by age, gender, education, ethnicity, and regional unit. Let $C$ index demographic cells with population counts $N_c$. In practice, this is estimated based on all the joint distributions provided publicly by the government statistical office.
- **Regional Turnout Data**: Official counts of actual voters $V_r$ for each regional unit $r$, representing the ground truth for aggregate turnout.
- **RDD Poll Data**: Random Digit Dialing survey conducted around election time with $N_{\text{poll}}$ respondents, recording both demographics and voting intention (Yes/No). This data suffers from selection bias as it is well known that survey response correlates with civic engagement (Abrajano et al. 2018; Sturgis and Jennings 2017).

The fundamental challenge is that the poll data, while rich in demographic information, cannot be assumed to be representative due to non-random response patterns. Traditional approaches require exclusion restrictions—variables affecting selection but not outcome—which are often theoretically contentious and empirically dubious. We replace this need by using ecological inference (King 1997; King et al. 1999) to ground the polling data in official statistics that cover the entire population and are thus entirely free from the selection effects.

## A.2 Model Specification

We specify a bivariate probit selection model where each eligible voter $i$ has latent propensities for selection (responding to survey) and outcome (turning out to vote) (Heckman 1978, 1979):

$$S_i^* = X_i\beta^s + \epsilon_i^s \tag{5}$$

$$O_i^* = X_i\beta^o + \epsilon_i^o \tag{6}$$

where $X_i$ is a $1 \times D$ vector containing (observed) demographic indicators for respondent $i$, $\beta^s$ and $\beta^o$ are both $D \times 1$ parameter vector variables, and the error terms $\epsilon^s$ and $\epsilon^o$ follow a bivariate normal distribution with correlation $\rho$:

$$\begin{pmatrix} \epsilon_i^s \\ \epsilon_i^o \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

and are assumed to be i.i.d with respect to $i$.

We also use binary indicators $S_i = \mathbb{I}(S_i^* > 0)$ for poll response and $O_i = \mathbb{I}(O_i^* > 0)$ for voting intention (for respondents).

## A.3 Integrated Likelihood Framework

The key innovation of our approach is incorporating ecological inference with a Heckman-style selection model for the joint modeling of all three data sources through a composite likelihood function.

In the following, when we talk about likelihood functions, we write them on the log-scale without normalizing constants, as used in MCMC sampling.

### Poll Selection Likelihood

Given the assumption of a completely representative initial contact set, the probability of observing a respondent from cell $c$ is proportional to the expected number of willing respondents:

$$\pi_c^s = \frac{\Phi(X_c\beta^s) \cdot N_c}{\sum_{c'} \Phi(X_{c'}\beta^s) \cdot N_{c'}}$$

The likelihood for the observed poll sample counts $\{n_c\}$ is then multinomial:

$$\mathcal{L}_{PS}(\beta^s) = \prod_c (\pi_c^s)^{n_c}$$

This identifies the selection model parameters $\beta^s$ up to scale, with relative response probabilities informed by the census population structure.

## Regional Turnout Likelihood

For each regional unit $r$, the probability of voting according to the model is given by the weighted average of the cells $c$ across the entire region $r$:

$$p_r = \frac{\sum_{c \in r} \Phi(X_c \beta^o) \cdot N_c}{\sum_{c \in r} N_c}$$

We model the official turnout counts $V_r$ as a binomial distribution:

$$\mathcal{L}_{RT}(\beta^o) = \prod_r p_r^{V_r} \left(1 - p_r\right)^{N_r - V_r}.$$

## Poll Response Likelihood

For poll respondents, we condition on selection ($S_i = 1$) and observe voting intention $O_i$ given the demographics vector $X_i$. Using the bivariate normal structure, their contribution to the likelihood is:

$$\begin{aligned}\mathcal{L}_{PR}(\beta^s, \beta^o, \rho) = &\prod_{\{i:O_i=1\}} \Phi_2(X_i\beta^o, X_i\beta^s; \rho) \\ &\times \prod_{\{i:O_i=0\}} \Phi_2(-X_i\beta^o, X_i\beta^s; -\rho) \\ &\times \prod_i \frac{1}{\Phi(X_i\beta^s)}\end{aligned}$$

where $\Phi_2$ denotes the bivariate normal CDF.

We can transform this to separate the effect of selection by applying data augmentation with latent variables

$$u_i \mid \beta^s \sim \text{TruncatedNormal}(-X_i\beta^s, \infty),$$

where $\text{TruncatedNormal}(b_l, b_u)$ is the standard normal distribution truncated from below at $b_l$ and above at $b_u$. This yields the conditional likelihood:

$$\mathcal{L}'_{PR}(\beta^o, \rho, u) = \prod_i \Phi\left(\frac{(2O_i - 1)(X_i\beta^o + \rho u_i)}{\sqrt{1 - \rho^2}}\right)$$

It is worth noting that fixing $\rho = 0$, this is the likelihood for simple Probit regression.

A more detailed description of the derivation is provided in Section A.6.

## A.4 Turnout models

We adopt a Bayesian framework, combining the likelihood components with prior distributions, and propose 4 distinct models:

**EI:** Ecological Inference model is just providing Regional Turnout likelihood $\mathcal{L}_{RT}$ along with priors on $\beta^o$:

$$p(\beta^o \mid \text{Data}) \propto \mathcal{L}_{RT}(\beta^o)p(\beta^o).$$

This is effectively a variation of the model in Rosen et al. 2001 with Beta-Binomial distribution replaced with a Probit and inputs generalized to multiple categories.

**BP:** Basic Probit regression model on survey data is achieved by setting $\rho = 0$ in the Poll Response likelihood $\mathcal{L}'_{PR}$ as noted above

$$p(\beta^s, \beta^o, \rho, u \mid \text{Data}) \propto \mathcal{L}'_{PR}(\beta^o, 0, \mathbf{0}) \cdot p(\beta^o),$$

**PM:** Poll and Margin model combines the two above while still ignoring the selection:

$$p(\beta^s, \beta^o, \rho, u \mid \text{Data}) \propto \mathcal{L}_{RT}(\beta^o) \cdot \mathcal{L}'_{PR}(\beta^o, 0, \mathbf{0}) \cdot p(\beta^o).$$

**FS:** Full Selection model finally introduces the Poll Selection component $\mathcal{L}_{PS}$:

$$p(\beta^s, \beta^o, \rho, u \mid \text{Data}) \propto \mathcal{L}_{PS}(\beta^s) \cdot \mathcal{L}_{RT}(\beta^o) \cdot \mathcal{L}'_{PR}(\beta^o, \rho, u) \cdot p(u \mid \beta^s)p(\beta^s)p(\beta^o)p(\rho),$$

**AFS:** Approximate Full Selection model simply approximates $u$ with the mean values of the truncated distributions provided by the Inverse Mills Ratio $\lambda(x) = \frac{\phi(x)}{\Phi(x)}$ so $u_i = \lambda(X_i^s \beta^s)$, which is a very common computational simplification with Heckman-style selection models. In our case, this drastically reduces the number of free variables and simplifies the geometry, at the potential cost of some accuracy.

All of the model formulations enable straightforward Hamiltonian Monte Carlo sampling while preserving the marginal posterior distribution of interest. Furthermore, as all four models infer a cell-level turnout probability, they can be used together with post-stratification to estimate the average for population and larger sub-groups.

## A.5 Multilevel Modeling

When dealing with categorical regressors, we can benefit from partial pooling over each category dimension with multilevel modeling. Suppose we have $D$ different category dimensions

(nationality, education, gender, age group etc) with $J[d]$ categories in each. We can then structure $X_i = X_{i,1}|\ldots|X_{i,D}$ as a concatenation of one-hot encoding vectors $X_{i,d}$ and similarly take $\beta = \beta_1|\ldots|\beta_D$. A simple two-level model for weights would then be defined by:

$$\beta_{d[j]} \sim \text{Normal}\big(0, \tau_d\big), \qquad \text{for } j = 1, \ldots, J[d] \tag{7}$$

$$\tau_d \sim \text{Half-Normal}(0, \sigma_0), \qquad \text{for } d = 1, \ldots, D \tag{8}$$

This allows the model to learn the expected effect size of each category, and helps properly regularize the less common categories to match the effect size of other categories in their magnitude. For log-likelihoods, this implies just replacing $p(\beta)$ with $\prod_d p(\beta_d|\tau_d)p(\tau_d)$.

## A.6 Derivation of Full Selection model

In this section we will perform a step-by-step derivation of the FS model.

### A.6.1 Bivariate Normal Properties

The bivariate normal cumulative distribution function $\Phi_2$ is defined as:

$$\Phi_2(x_1, x_2; \rho) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \phi_2(u, v; \rho)dvdu$$

where $\phi_2$ is the bivariate normal probability density function:

$$\phi_2(u, v; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(u^2 - 2\rho uv + v^2\right)\right\}.$$

A key identity for computational purposes expresses the bivariate CDF in terms of univariate integrals:

$$\Phi_2(x_1, x_2; \rho) = \int_{-x_2}^{\infty} \Phi\left(\frac{x_1 + \rho u}{\sqrt{1-\rho^2}}\right)\phi(u)du \tag{9}$$

where $\phi(\cdot)$ is the standard normal PDF and $\Phi(\cdot)$ is the standard normal CDF.

### A.6.2 Standard Bivariate Probit Likelihood

The complete-data likelihood for the bivariate probit selection model is:

$$\begin{aligned}\mathcal{L}(\beta^o, \beta^s, \rho) = &\prod_{\{i:S_i=0\}} \Phi\left(-X_i\beta^s\right) \\ &\times \prod_{\{i:S_i=1, O_i=1\}} \Phi_2\left(X_i\beta^o, X_i\beta^s, \rho\right) \\ &\times \prod_{\{i:S_i=1, O_i=0\}} \Phi_2\left(-X_i\beta^o, X_i\beta^s, -\rho\right),\end{aligned}$$

where $S_i$ indicates selection (response to poll) and $O_i$ indicates turnout intention.

### A.6.3 Data Augmentation Representation

We introduce latent variables $u_i \sim \mathcal{N}(0,1)$ for each respondent to demarginalize the bivariate normal distribution. Using identity (9), we obtain the augmented likelihood:

$$\begin{aligned}\mathcal{L}_{\text{aug}}(\beta^o, \beta^s, \rho, \{u_i\}) = &\prod_{\{i:S_i=0\}} \Phi\left(-X_i\beta^s\right) \\ &\times \prod_{\{i:S_i=1,O_i=1\}} \Phi\left(\frac{X_i\beta^o + \rho u_i}{\sqrt{1-\rho^2}}\right) \\ &\times \prod_{\{i:S_i=1,O_i=0\}} \Phi\left(\frac{-(X_i\beta^o + \rho u_i)}{\sqrt{1-\rho^2}}\right) \\ &\times \prod_{\{i:S_i=1\}} \mathbb{I}[u_i > -X_i^s\beta^s]\phi(u_i).\end{aligned}$$

such that marginalizing over $\{u_i\}$ recovers the original likelihood:

$$\int \mathcal{L}_{\text{aug}}(\beta^o, \beta^s, \rho, \{u_i\}) d\{u_i\} = \mathcal{L}(\beta^o, \beta^s, \rho).$$

### A.6.4 Conditioning on Selection

In our case, we are interested in the joint model only for the polling data for which $S_i = 1$. We therefore need to condition on that being the case for all $i$, which just means dividing by $P(\forall i : S_i = 1) = \prod_i \Phi(X_i^s\beta^s)$, yielding

$$\begin{aligned}\mathcal{L}_{\text{aug}}(\beta^o, \beta^s, \rho, u \mid \{S_i = 1\}) = &\prod_{\{i\}} \Phi\left(\frac{(2O_i - 1)(X_i\beta^o + \rho u_i)}{\sqrt{1-\rho^2}}\right) \\ &\times \prod_{\{i\}} \mathbb{I}[u_i > -X_i^s\beta^s]\frac{\phi(u_i)}{\Phi(X_i^s\beta^s)}.\end{aligned}$$

Considering $\mathbb{I}[u_i > -X_i^s\beta^s]\frac{\phi(u_i)}{\Phi(X_i^s\beta^s)}$ is the density function for Truncated Normal, this is equivalent to just changing the prior for $u_i$ to a much cleaner

$$u_i \sim \text{TruncatedNormal}(-X_i^s\beta^s, \infty).$$

This formulation enables efficient HMC sampling while maintaining the exact bivariate normal structure of the selection model.

# License

## Non-exclusive licence to reproduce thesis and make thesis public

I, **Reimo Rebane**,

*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   **Exploring a novel estimation method for voter turnout**,

   *(*title of thesis)

   supervised by Margus Niitsoo, Tarmo Jüristo and Andres Võrk.

   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Reimo Rebane

***15/01/2026***