SOEP-Core – 2021: Sampling, Non-response, and Weighting in Wave 2 of Living in Germany – Nationwide Corona-Monitoring (RKI-SOEP2)

Christian Danne, Maximilian Priem, Hans Walter Steinhauer

DIW S●EP

Running since 1984, the German Socio-Economic Panel (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:
Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at http://www.diw.de/soepsurveypapers

Please cite this paper as follows:

Christian Danne, Maximilian Priem, Hans Walter Steinhauer. 2022. SOEP-Core – 2021: Sampling, Nonresponse, and Weighting in Wave 2 of Living in Germany – Nationwide Corona-Monitoring (RKI-SOEP2). SOEP Survey Papers 1175: Series C. Berlin: DIW/SOEP

Sampling, Nonresponse, and Weighting in Wave 2 of Living in Germany – Corona-Monitoring (RKI-SOEP2)

Christian Danne[1], Maximilian Priem[*1], and
Hans Walter Steinhauer[2]

[1]DIW Econ GmbH
[2]DIW Berlin

Last updated: July 29, 2022

This paper describes the weighting methodology for the second wave of the Living in Germany - Nationwide Corona Monitoring study (RKI-SOEP2). Information from the larger, ongoing German Socio-Economic Panel (SOEP) study from which the households are sampled is used to analyze the determinants of non-contact, attrition, and refusal to participate at the household and individual level as well as to correct the data accordingly. For a sample of 6,791 households and 11,223 individuals, we find that personal attributes and socio-economic factors at the household level and the advent of the omicron variant are the key determinants at the household level for non-response. Demographic and socio-economic factors represent the most important predictors for non-response at the individual level.

JEL Classifications: H12, I10, I18, C83.
Keywords: Calibration, Weighting, Covid-19, SOEP, RKI, Lasso.

*Corresponding author: Maximilian Priem, DIW Econ GmbH, Mohrenstrasse 58, 10117 Berlin, mpriem@diw-econ.de

1 Introduction

Determining the exact SARS-CoV-2 seroprevalence in the general population is key for designing appropriate measures to combat the Covid-19 pandemic and *ex-post* evaluations of these policies. So far, policy decisions in Germany are largely based on Covid-19 test data collected by local health authorities and published by the Robert Koch Institute (RKI).[1]

A number of papers argue that the published data is likely to underestimate the true SARS-CoV-2 prevalence due to selectivity in Covid-19 testing (cf. Rendtel et al. 2020). To a large extent, Covid-19 tests are administered to persons already showing signs of a potential SARS-CoV-2 infection. On the other hand, the rise of SARS-CoV-2 cases with no or only mild symptoms due to the increased prevalence of the Omicron variant and limited testing facilities implies that many SARS-CoV-2 cases go unnoticed.

In order to determine the true extent of the SARS-CoV-2 seroprevalence, the German Socio-Economic Panel (SOEP) at the German Institute for Economic Research (DIW Berlin) and the RKI launched the Living in Germany – Nationwide Corona Monitoring study (RKI-SOEP) in October 2020. For this seroepidemiological study, households were selected from the larger, ongoing SOEP survey. Households and household members were surveyed about their past and present Covid-19 infections, possible symptoms, recent Covid-19 testing, and a number of general health behaviors in relation to the Covid-19 pandemic. In addition, participants were asked to provide a blood sample for a lab-based Immunoglobulin G antibody-test in order to detect a previous Covid-19 infection along with their survey responses.

Efficient and unbiased estimation of the true seroprevalence requires adjusting the collected data for a potential non-response bias at the household and individual level as well as for over- or under-representation of certain groups in the sample.[2]

This paper describes the weighting methodology for the second wave of the RKI-SOEP study (RKI-SOEP2). The approach largely follows Siegers, Belcheva, and Silbermann (2020) and the re-weighting procedure for the first wave of the RKI-SOEP survey as described in Steinhauer, Zinn, and Siegers (2021). The procedure utilizes data from previous year's waves taken from the SOEP to analyze non-contact, attrition, and refusal. The fact that households are selected from the SOEP implies that we have the same set of information for each household and its members, regardless of whether the household participated in the RKI-SOEP2 study or not. Moreover, the SOEP data provides a vast set of variables that can be used to analyze and to adjust the non-response bias present in the data.

We start with more than 500 variables at the district, household, and individual levels and use a Lasso approach to select the model specification for non-contact, attrition, and refusal. The predictions from these models are used to re-weight the pre-existing household and individual weights taken from the last wave of the underlying SOEP wave to generate the weights for the RKI-SOEP2 study. In a last step, we adjust the resulting weights for over- and under-representation of certain sub-groups in the population.

---

[1] See https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4.

[2] Since the SOEP study samples at the household level, non-response bias can occur at the household and individual levels.

For a sample of 6,791 households and 11,223 individuals, the results show that personal traits and socio-economic factors at the household level as well as the advent of the omicron variant are the key determinants at household level for non-response. While, demographic and socio-economic factors represent the most important predictors for non-response at the individual level. We use those factors to construct longitudinal weights at the household and individual levels for both the entire RKI-SOEP2 sample as well as for refugee and migrant sub-populations.

The remainder of the paper is as follows. Section 2 describes the SOEP-RKI2 data, as well as the underlying SOEP study and the sampling of the SOEP-RKI2 survey. Section 3 describes the weighting methodology. Section 4 describes the variable selection algorithm. Sections 5 and 6 present the results of the final models at the household and individual levels, as well as the trimming and weighting procedures. 7 extends the procedure to two subsets of the study, namely the sample that was surveyed prior to omicron becoming the predominant Covid-19 variant and the *core* sample of the SOEP. Section 8 concludes.

2 Data, Sampling, and Stratification

The SOEP longitudinal household survey was launched in 1984 with the aim to collect representative data on socio-economic, physical, and mental health variables as well as attitudes, values, and data on personality traits at the household and individual levels. Every year, approximately 15,000 households and 30,000 individuals are surveyed by the SOEP's fieldwork organization. The data consists of representative (sub-) samples of the population and several sub-groups across Germany. In addition to the *core* SOEP sample, several enlargement samples exist in order to address panel attrition and capture changes in the general population due to immigration (samples M1 and M2 Liebig et al. (2021) and refresher samples M7 and M8) or an increase in the number of refugees (samples M3-M6). These samples are typically integrated after periods of increased gross influx of either of the two groups.[3] For the remainder of the paper, the combined data set consisting of the *core* sample and the M1-M8 samples is referred to as "SOEP" data.

In October 2020, SOEP researchers, jointly with the RKI, launched the first wave of the RKI-SOEP Corona monitoring study in order to study the SARS-CoV-2 seroprevalence among the German population. Households were sampled from the SOEP population. Participation in the survey was voluntary. Individuals in each household were interviewed on past and present Covid-19 infections, symptoms, recent Covid-test results, and general health questions in relation the Covid-19 pandemic. In addition, participants were asked to submit a self-taken blood sample for a lab-based antibody test along with their survey responses.

The RKI-SOEP2 survey started in November 2021. Table 1 shows the sample sizes and participants over the course of the fieldwork period between November 2021 and March 2022. While the first wave of RKI-SOEP was restricted to household members 18 years

---

[3]Migrant and refugee samples are joint projects of the SOEP, the Federal Office for Migration and Refugees (BAMF), and the Institute for Employment Research's (IAB) research data center. See Goebel et al. (2019) and Nebelin, Petrenz, and Wenzig (2019) for a description of the samples.

or older, the second wave also allowed children aged 14 years and older to participate. Between November 2021 and March 2022, a total of 6,791 households and 11,223 individuals were surveyed in three different tranches.

Tranching of households was based on the progress of the fieldwork of the main SOEP survey. The aim was to achieve a certain time gap between participation in the SOEP survey and the RKI-SOEP2 study. Thus, households that were surveyed early in the year for the main study entered the first tranche. Households that were surveyed towards the middle and the end of the SOEP fieldwork period, entered tranches 2 and 3 respectively. Fieldwork ended on March, 6th 2022. A detailed description of the fieldwork can be found in Hess and Torregroza (2022).

Table 1: Sample Sizes and Participants Over the Course of the Fieldwork Period

	Fieldwork	Sample		
Tranche	Month	HHs	Individuals	Participants
1	November 2022	7,697	12,729	7,906
2	December 2022	1,191	2,079	813
3	January 2022	3,213	6,648	2,504
Total		**12,101**	**21,456**	**11,223**

The survey was initially announced to 12,101 households part of the SOEP by a letter from the German Ministry of Health accompanied by additional information about the survey and request to participate. Shortly after, the official recruitment letter from the field service provider was sent out, containing the questionnaire, a privacy policy statement, the declaration of consent to be signed by the household members, the participation plan, and a dry blood sample (DBS) test kit. Test kits were accompanied by detailed instructions on how to self-administer the test as well as professional package material to ensure safe return-shipment of the test samples. Hoebel et al. (2021) provide a detailed overview of the questionnaire, the test kits, and additional material provided to households. Households who did not respond to the initial recruitment letter within two weeks subsequently received a reminder.

Figure 1 summarizes households and individuals according to their decision to participate or not in the survey. The recruitment efforts yielded 6,791 households and 11,223 individuals who were willing to participate in the study and submitted a signed form of consent. Of these 11,223 household members, 10,524 persons submitted a questionnaire, 10,687 provided a valid DBS. Thus, 8,586 individuals could not be reached. Another 663 individuals had moved and their new addresses could not be determined. After being contacted by the interviewer, 965 individuals refused participation. Further, 6 individuals had moved abroad since their last participation in the SOEP survey. Lastly, 13 individuals were deceased.
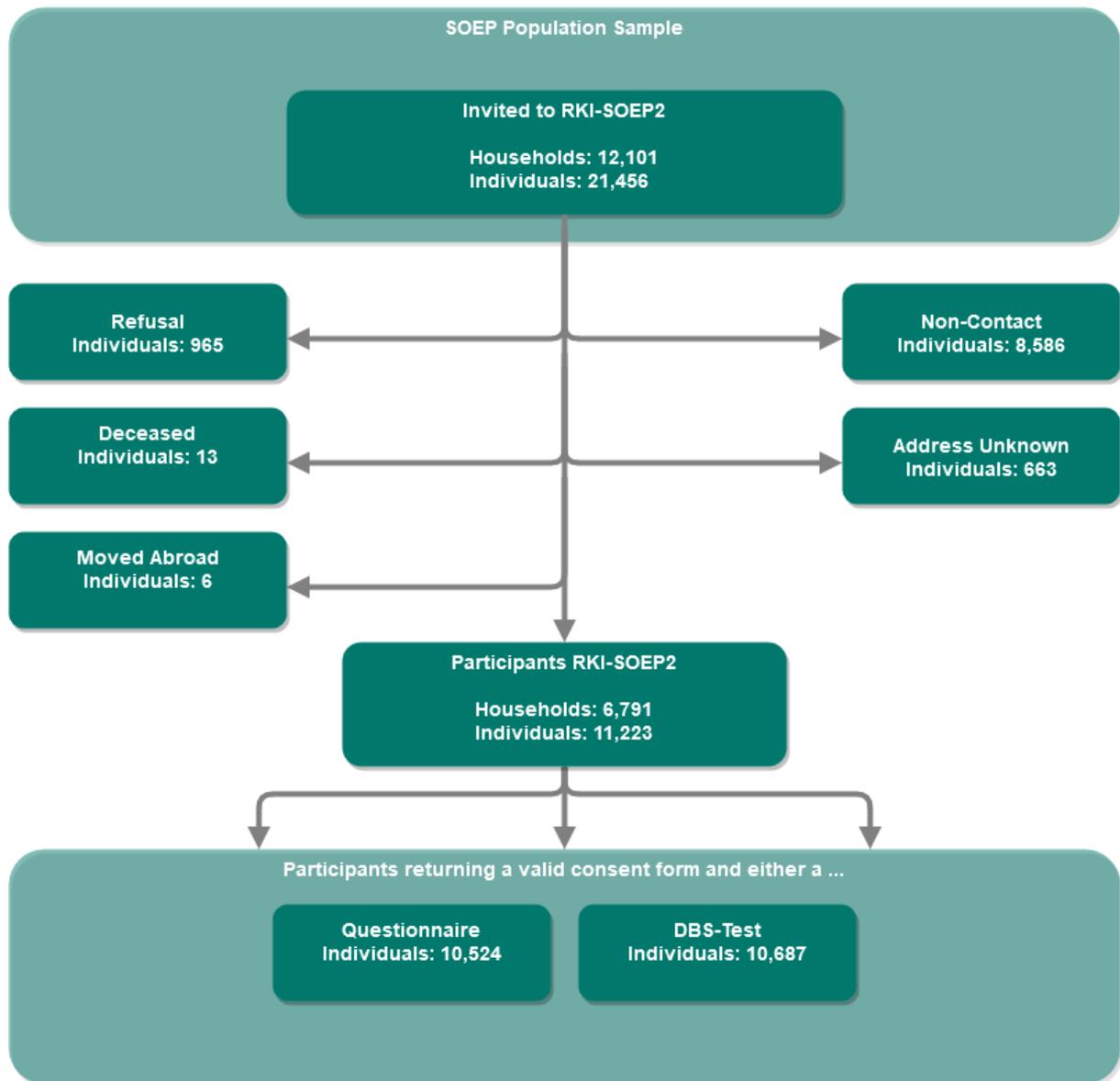
Figure 1: Participation vs. Non-Contact, Attrition, and Refusal

Figure 2 shows the influx of test kits and questionnaires over the course of the field period. As shown by the graph, the majority of items were returned in November and early December 2021 as well as in the second half of January 2022.
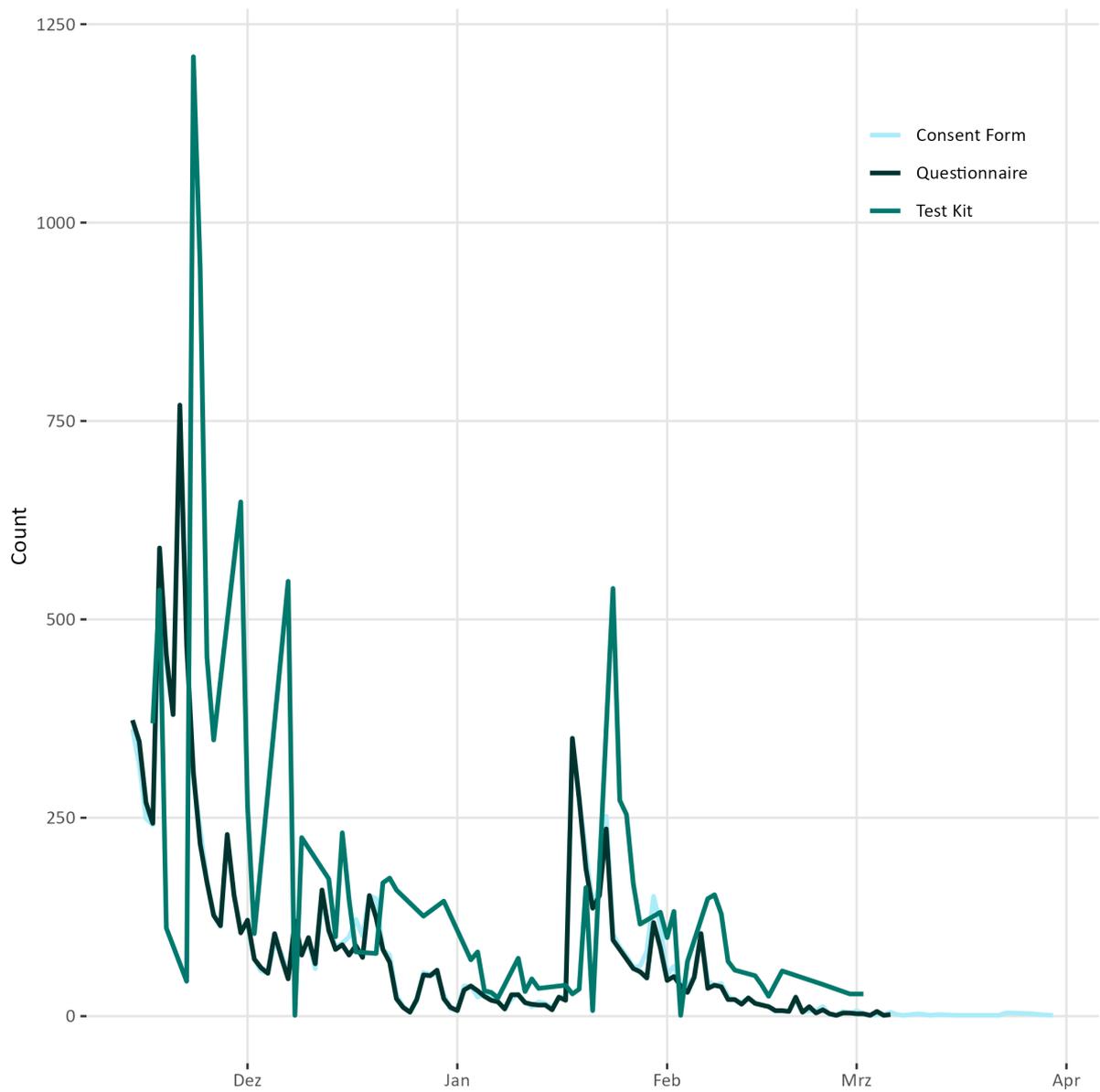
Figure 2: Return of Questionnaires, Consent Forms, and Test Kits

Figure 3 shows the composition of the SOEP sample by sex and age of participating individuals. Although slightly skewed toward females, the distribution of the sample data matches the distribution of the underlying SOEP data.
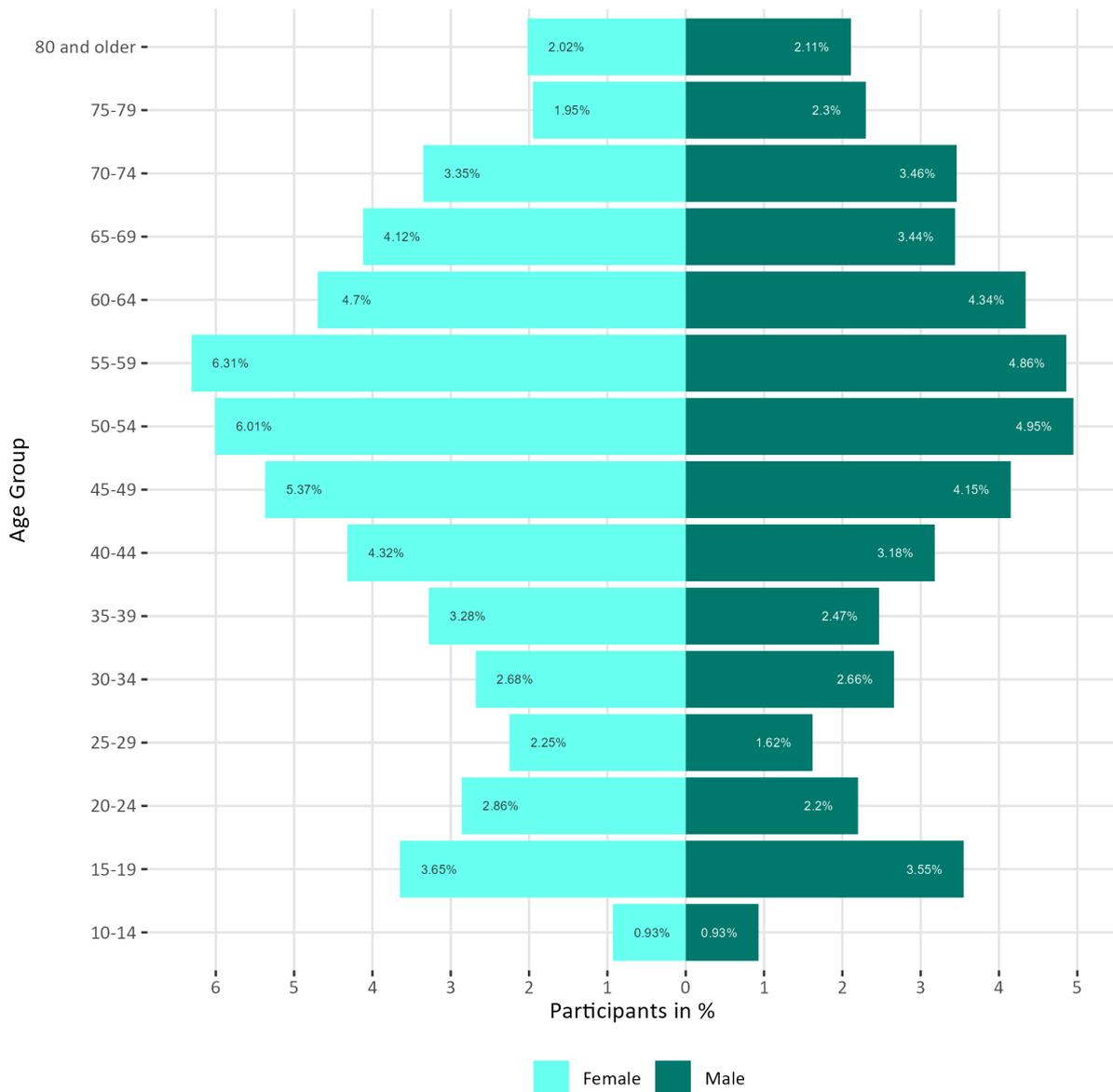
Figure 3: Individuals in the RKI-SOEP2 by Sex and Age

3 Weighting Methodology

The method for generating survey weights for the resulting sample largely follows Siegers, Belcheva, and Silbermann (2020) and Steinhauer, Zinn, and Siegers (2021). Since the RKI-SOEP2 study was conducted between two SOEP survey waves and design factors are already contained in the pre-existing SOEP panel weights, we use the most recent weight for each household as base weights and adjust these for those factors driving non-response. For most households and individuals in the sample, this corresponds to the weight of the 2020 SOEP survey year.

We define refusal as *ex-ante* declining to partake and remain in the RKI-SOEP survey. Non-participation is defined as not submitting a complete set of documents, i.e. the test

kit, questionnaire, and accompanying signed consent form. These two decisions are separately measured for each household and household members. At the individual level, we additionally correct for selectivity in self-testing, since the ability to take a blood sample without assistance is likely to drive non-participation.

The resulting factors are then used to adjust the existing household and person weights from the SOEP. The resulting household weights are trimmed and adjusted for over and under-representation using the margins for household size, federal state, household type, and whether or not the household is an owner occupant. Individual weights are adjusted for age, sex, and migrant status. We use the raking procedure by DeVille, Särndal, and Sautory (1993) to carry out all adjustments.

4 Data and Model Selection

The starting point for modeling the non-response decision processes is a set of more than 500 explanatory variables at the district, neighborhood, household, and individual levels. Since a large number of the explanatory variables can be assumed to be either irrelevant or to be highly correlated with each other, we first pre-select a set of candidate variables for the final models to ease the computational burden and then use a Lasso model to find the final specifications for the models in order to avoid over-fitting.

The procedure is as follows. First, bivariate models are run for all possible combinations of dependent and independent variables and only those variables significant at the 1%-level are kept. Complimentary log-log regressions (cloglog) are used for this first step in order to account for the skewed distributions of the variables. Second, we calculate bivariate correlations for all remaining explanatory variables for each model. If the absolute value of the correlation coefficient between two independent variables exceeds 0.90, the variable with the smaller coefficient in the corresponding bivariate model is dropped from the set. All other variables remain in the model. Third, we estimate cloglog model with a Lasso shrinkage as suggested by Tibshirani (1996) and Chen et al. (2018) in order to get the final model for predicting the non-response probabilities. We use a 10-fold cross validation where the tuning-parameter λ is picked, such that the mean cross-validated error is minimized.

The data sources for the models are as follows. Household and individual data on demographics, health behavior, education, family composition, finances, personality traits, migration background, and political attitudes are taken from the SOEP data set. Socio-economic variables at the neighborhood level are taken from Microm and the Inkar database of the Federal Office for Building and Regional Planning.[4] Information about the current Covid-19 infection rates and cumulative incidence at the district level on the day the test sample kit was shipped to the household is taken from the RKI Covid-19 data hub.

---

[4]see https://www.microm.de and https://www.inkar.de/

5 Household-Level Results

The resulting models contain between 17 and 113 variables. Figure 4 shows the coefficient plot for the household's decision to not *ex-ante* refuse to partake in the survey. Since the model selection algorithm above only generates models with "significant" coefficients by construction, standard errors are dropped for convenience. As shown by Figure 4, the majority of remaining predictors in the model belong to the group of socio-economic factors at the household level and to the group of personal attributes of the (head of the) household, followed by demographic, geographic, and economic variables at the district level. In addition, the onset of the omicron variant had the strongest impact on the likelihood to remain in the sample.[5]

---

[5]We assume that prior to 2022, delta was the predominant variant of the virus. Thereafter, omicron was the most common variant.

Note: Dependent variable: y = 1, iff household remains in the RKI-SOEP survey and zero otherwise.
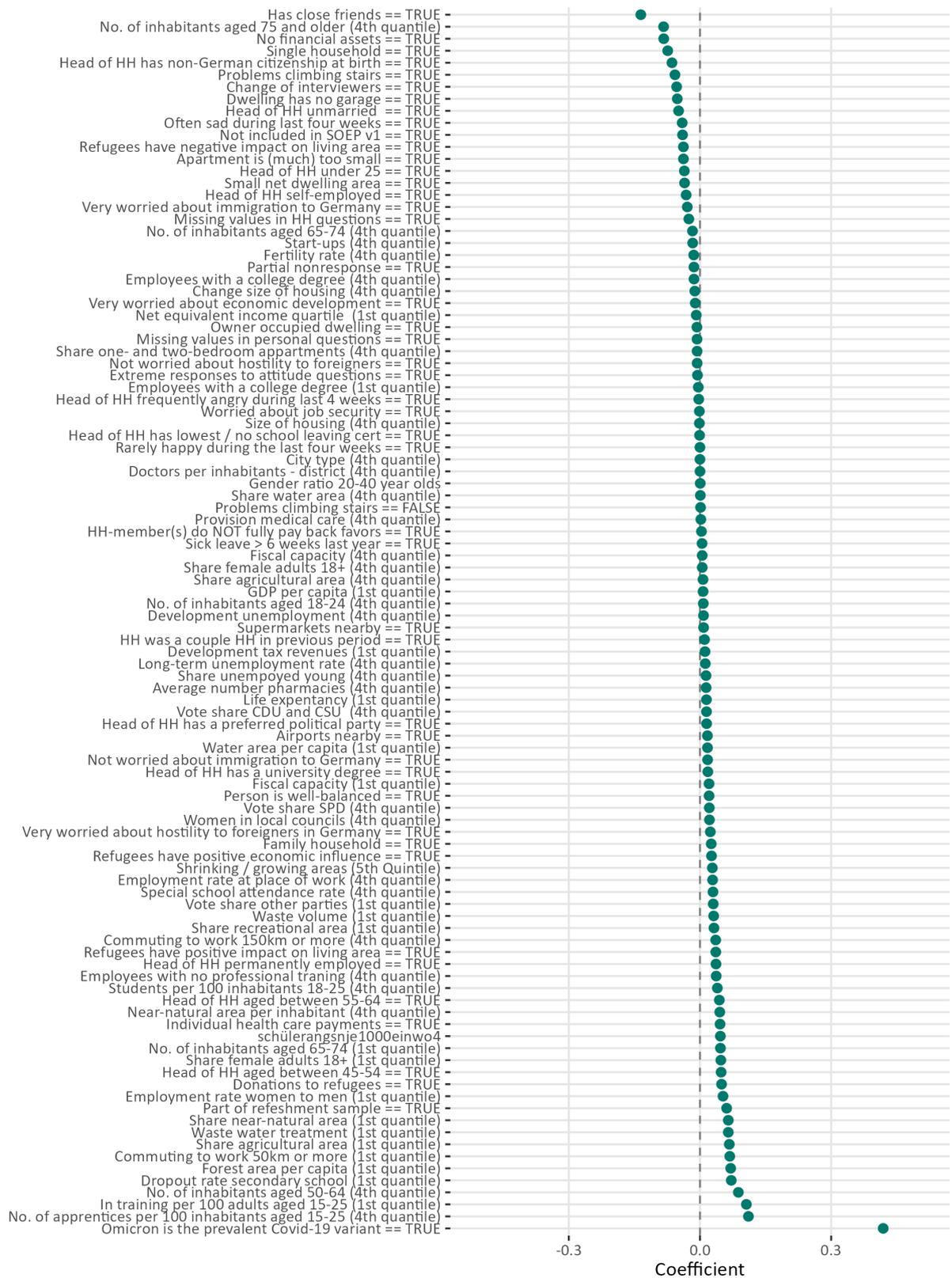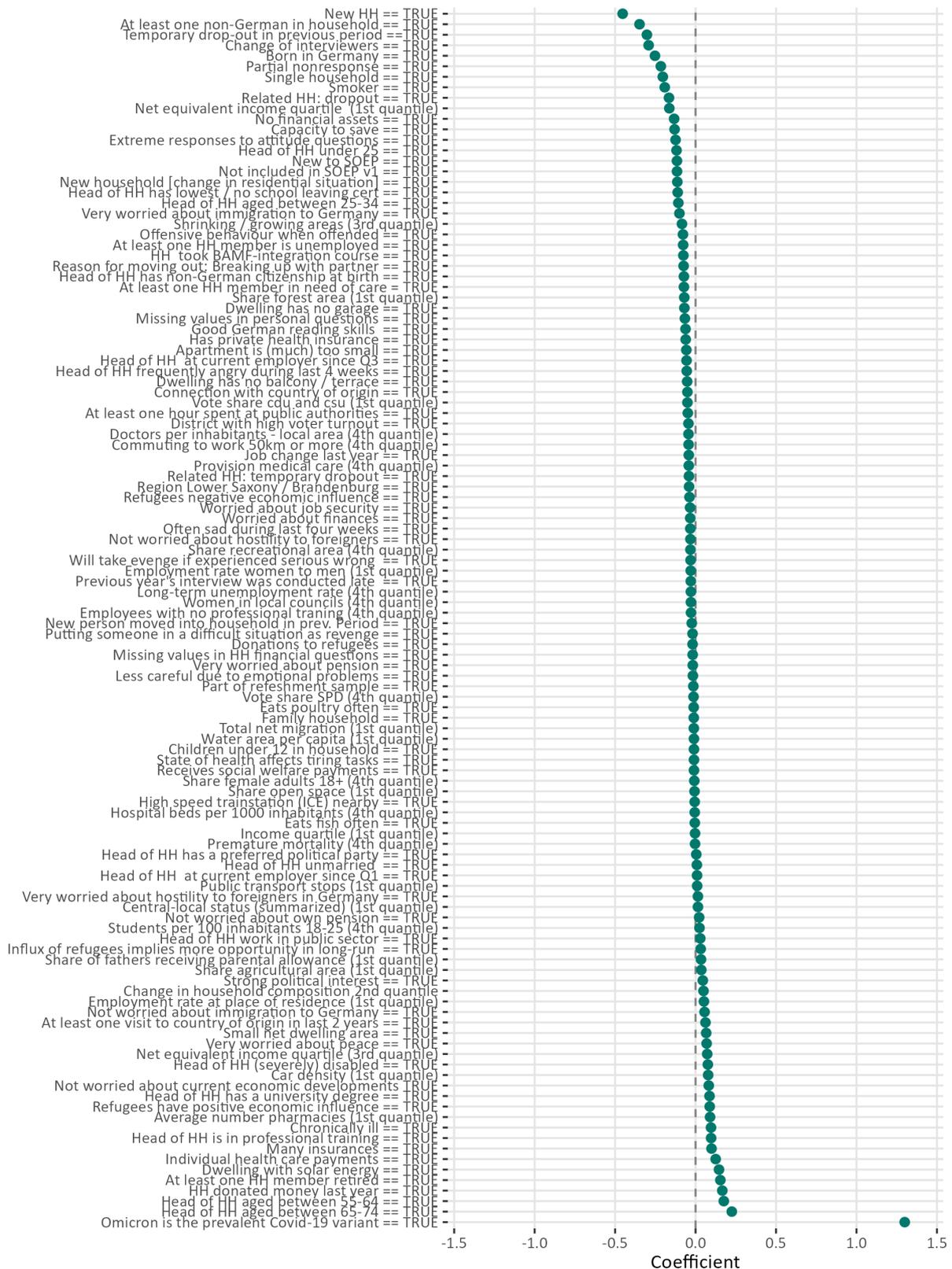Standard errors and the coefficient of the model's constant are omitted.

Figure 4: Household's Decision to Remain in the RKI-SOEP Survey

Figure 5 shows the coefficient plot for the household's decision to participate in the survey. While the variables themselves differ, they do belong to the same groups as the variables in the previous model. However, survey-related variables, such as whether or not households are new to the SOEP sample and variables related to a refugee or migratory status are also important.

Note: Dependent variable: y = 1, iff household participates in the RKI-SOEP survey and zero otherwise. Standard errors and the coefficient of the model's constant are omitted.

Figure 5: Household's Decision to Participate in the RKI-SOEP Survey

As outlined above, the predicted values from each of the models are taken to adjust the pre-existing household adjustment factors, such that the new adjustment factor for each household ($F_{hh,t}$) is given by

$$F_{hh,t} = F_{hh,soep,t-1} \cdot \frac{D_{r,t}}{\widehat{P}_{r,t}} \cdot \frac{D_{p,t}}{\widehat{P}_{p,t}} \qquad (1)$$

where $F_{hh,soep,t-1}$ is the household's adjustment factor taken from the last SOEP wave, $D_{r,t}$ and $D_{p,t}$ are dummy variables that are 1 if the household remains (r) in the survey and if the household participates (p) in the survey and zero otherwise. $\widehat{P}_t$ represents the corresponding predicted probabilities for r and p from the model.

The resulting household factors are trimmed and adjusted for over- and under-representation using margins for household size, federal state, household type, and whether or not the household is an owner occupant using the raking procedure by DeVille, Särndal, and Sautory (1993).

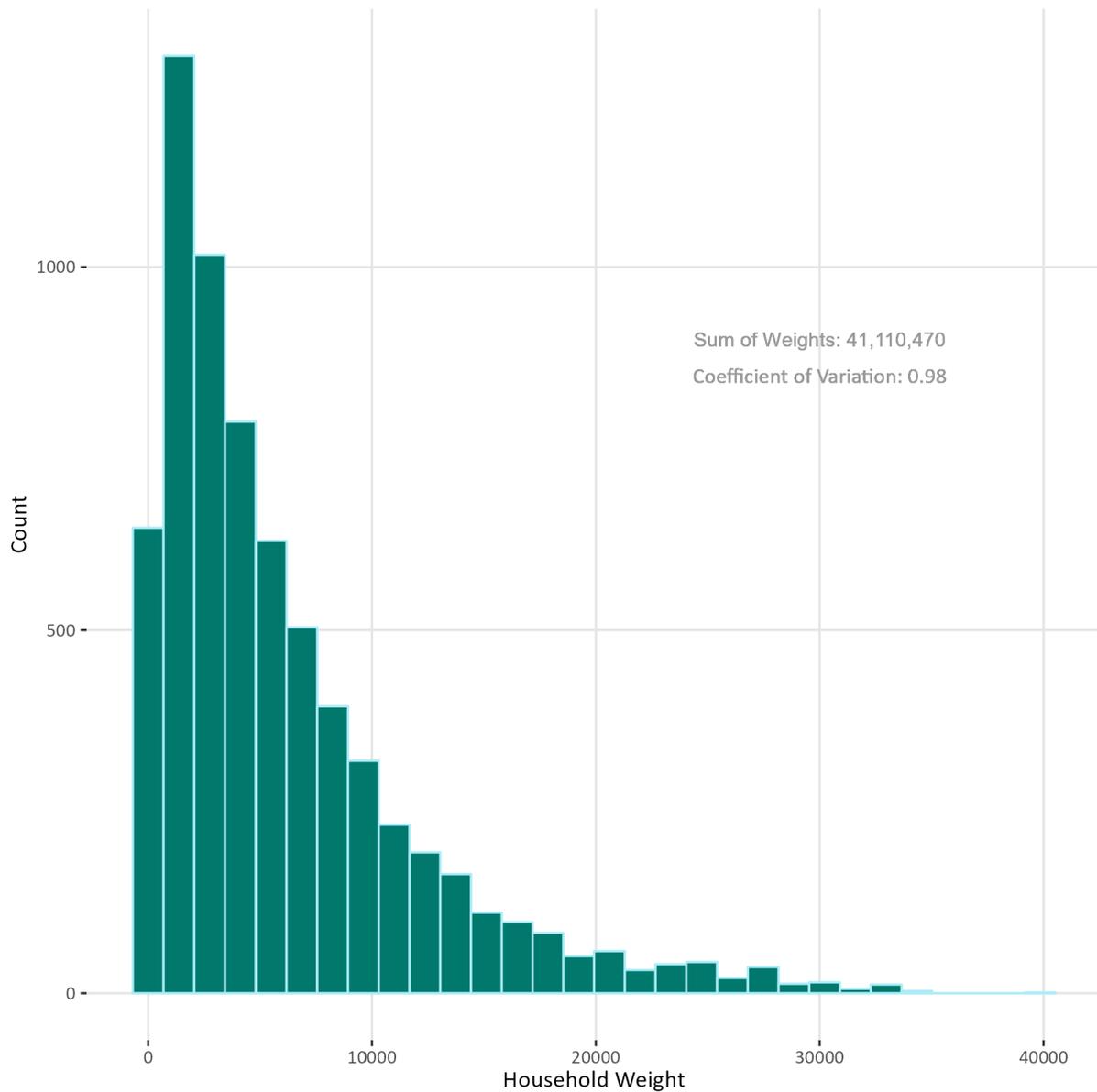Figure 6 shows the distribution of the resulting household weights as a reference.

Figure 6: Distribution of RKI-SOEP2 Household Weights

6 Individual-Level Results

Figure 7 shows the coefficient plot for the individual's decision to remain in the RKI-SOEP2 study. A lot fewer predictors are important at the individual level and can be largely described by age and variables related to the household's dwelling.
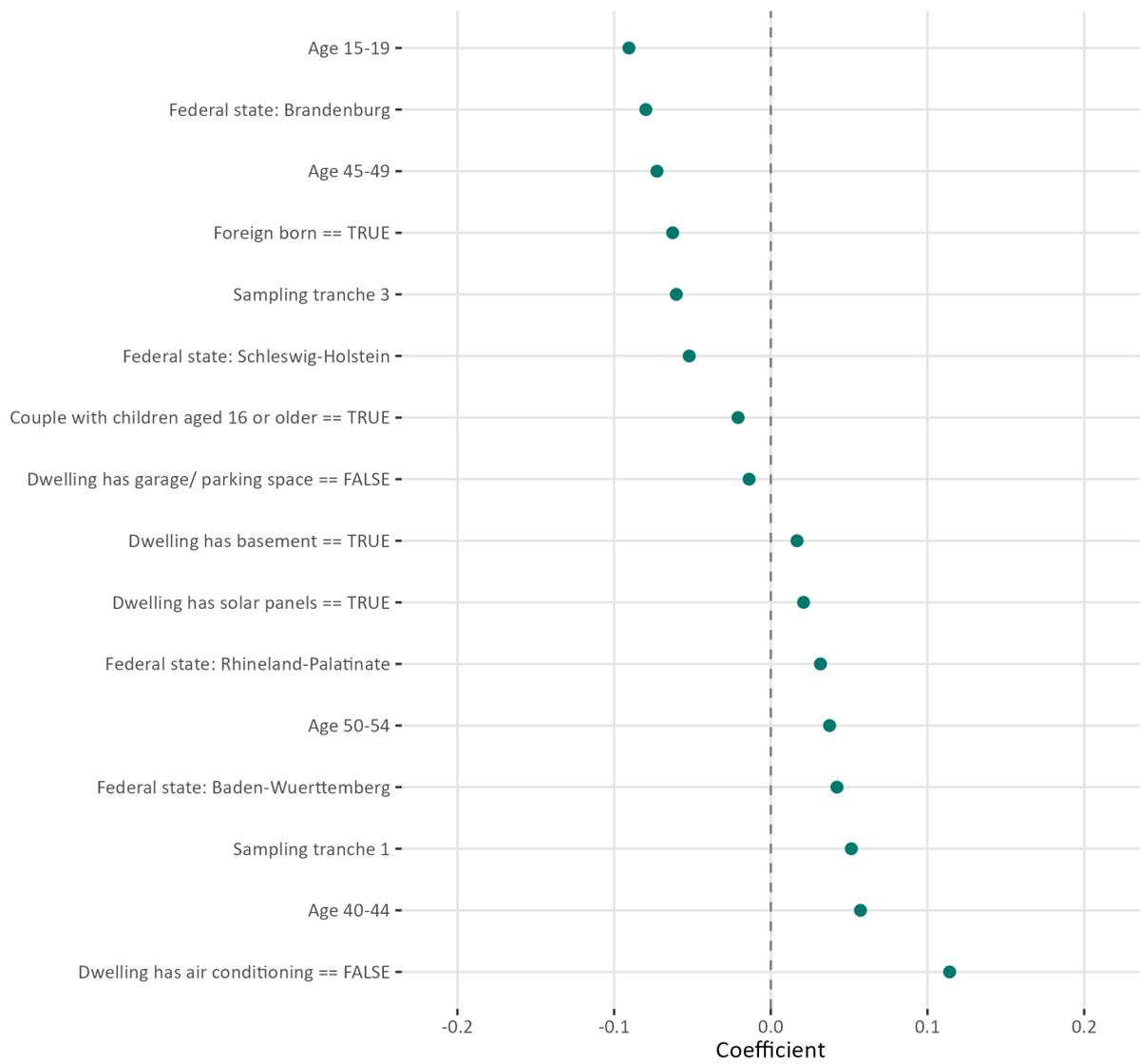
Note: Dependent variable: y = 1, iff individual remains in the RKI-SOEP survey and zero otherwise.
Standard errors and the coefficient of the model's constant are omitted.

Figure 7: Individual's Decision to Remain in the RKI-SOEP Survey

Figure 8 shows the coefficient plot for the individual's decision to participate in the RKI-SOEP2 study. In addition to age and attributes of the household's dwelling, additional household and individual traits help to predict the probability of taking part in the survey.
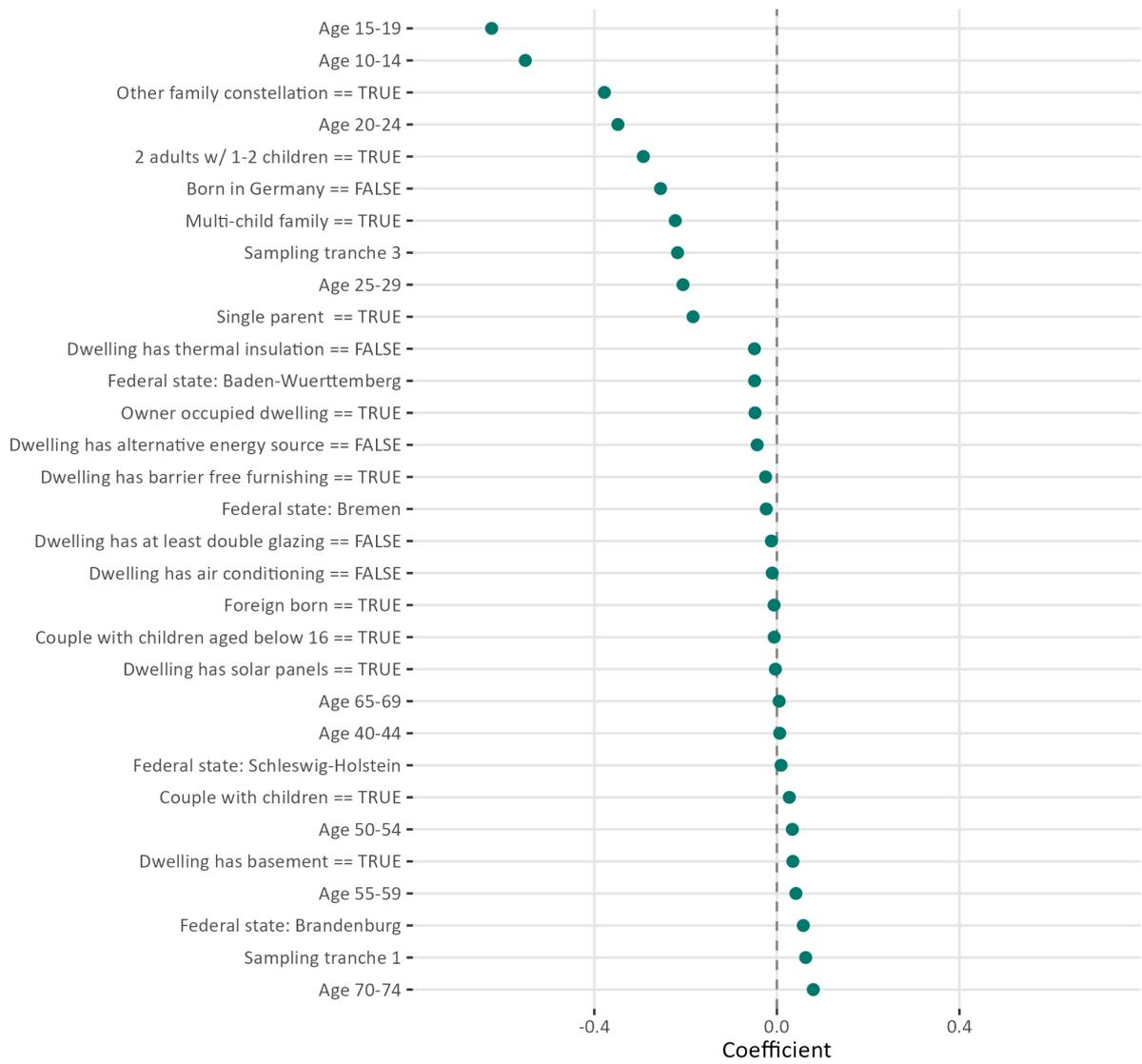
Note: Dependent variable: y = 1, iff individual participates in the RKI-SOEP survey and zero otherwise. Standard errors and the coefficient of the model's constant are omitted.

Figure 8: Individual's Decision to Participate in the RKI-SOEP Survey

Figure 9 shows the coefficient plot for the individual's decision to provide a valid DBS. The results largely overlap with those in Figure 8. However, age appears to play a much bigger role.
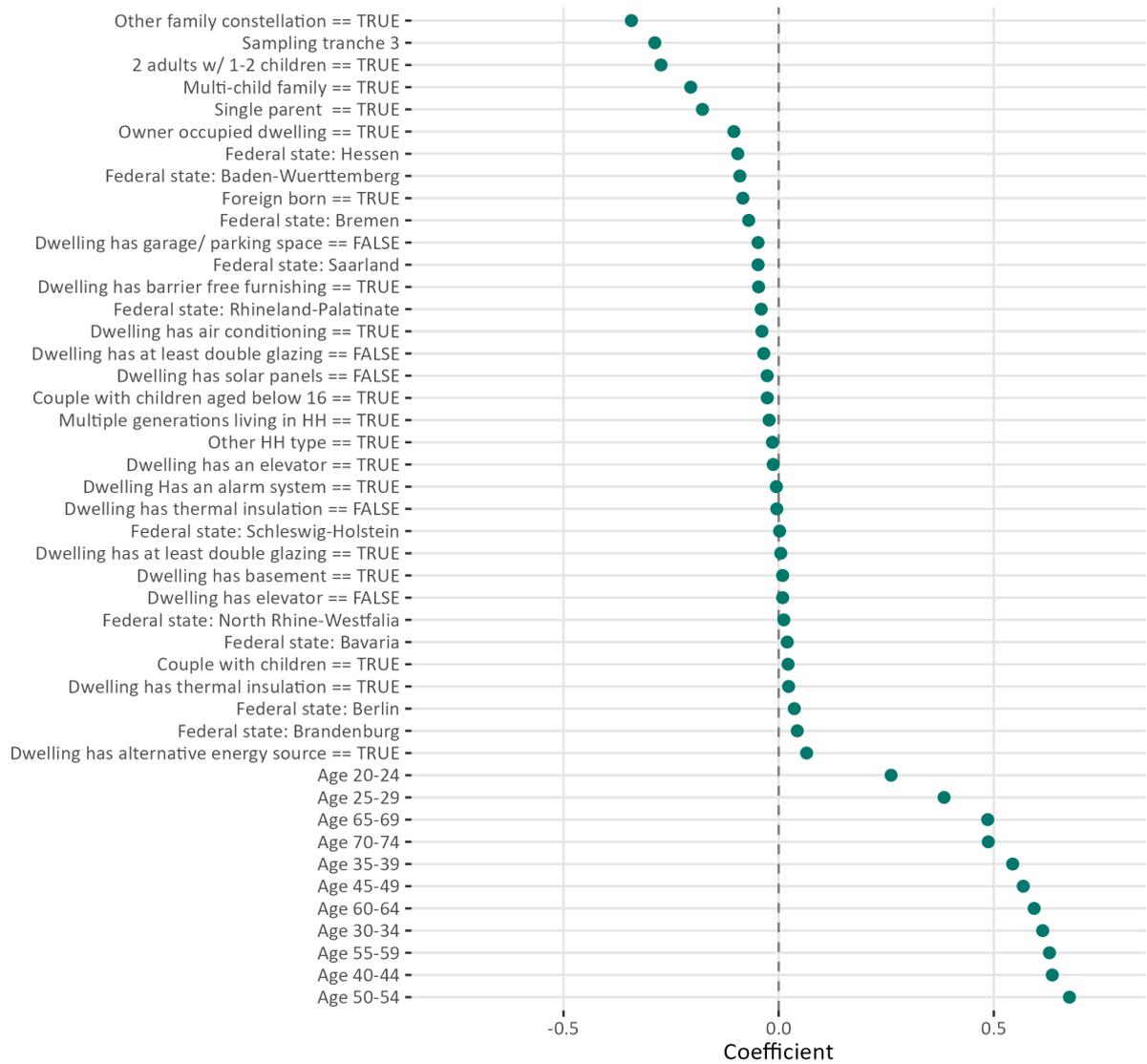
Note: Dependent variable: y = 1, iff individual provides a testable dry blood sample in the RKI-SOEP survey and zero otherwise. Standard errors and the coefficient of the model's constant are omitted.

Figure 9: Individual's Decision to Provide a Testable Dry Blood Sample

To calculate the individual weights, we start off using re-calculated household weights from the previous section and equally distribute them over the household members. We then use the predicted values from each of the models to adjust these generated starting weights, such that the new adjustment factor for each individual ($F_{p,t}$) in the sample is given by

$$F_{p,t} = \widehat{F}_{p,soep,t-1} \cdot \frac{D_{r,t}}{\widehat{P}_{r,t}} \cdot \frac{D_{p,t}}{\widehat{P}_{p,t}} \tag{2}$$

where $F_{p,soep,t-1}$ is the fraction of the generated household weight for each household member, $D_{r,t}$ and $D_{p,t}$ are dummy variables that are 1 if the individual remains (r) in the survey and if the household member participates or has submitted a testable blood sample

(p), respectively; zero otherwise. $\widehat{P}_t$ represents the corresponding predicted probabilities for r and p from the model.

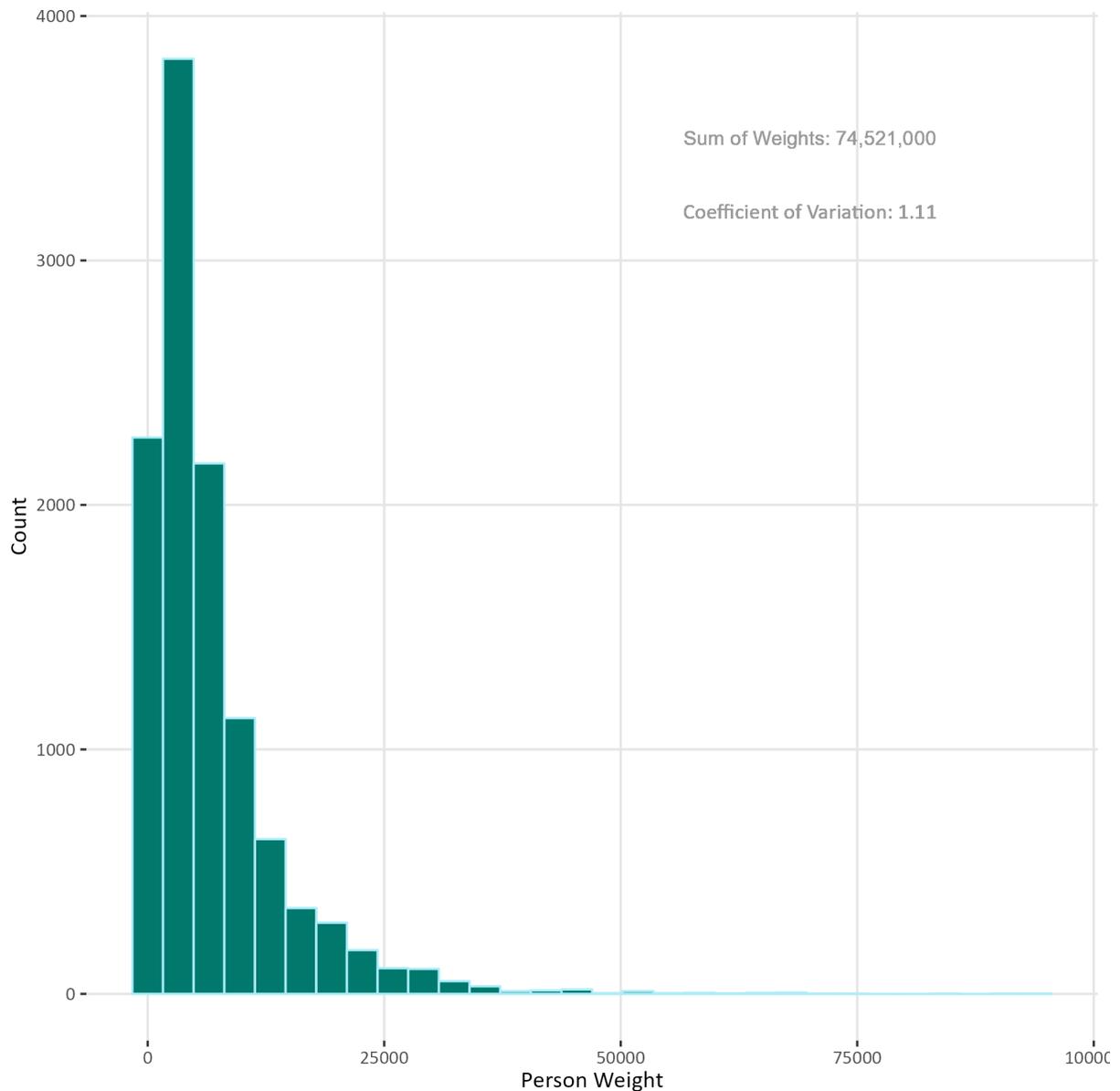Figure 10 shows the distribution of the resulting individual weights for reference.

Figure 10: Distribution of RKI-SOEP2 Individual Weights

7 Extensions

We use the same methodology as above to generate weights for two subsets of the study. First, weights are re-estimated for the sub-sample of households that were surveyed prior to omicron becoming the most prevalent variant of the virus. Second, we re-estimate the weights for the actual *core* sample of the study, namely by dropping individuals from the refugee samples M3-M6 described in section 2. We assume that prior to 2022, delta was

the main cause of a Covid-19 infection. Whereas from 2022 onwards, omicron was the main culprit.

Figures 11 and 12 show the coefficient estimates for households to remain in the survey and to participate in the survey, respectively. In both cases a much wider array of topics help to predict the probability to remain and to partake in the survey, when compared to the results above. Personal traits appear to be the most important factors, followed by socio-economic indicators at the household and individual levels. Moreover, the decision to remain in the survey prior to onset of omicron is associated with considerably fewer predictors than for the entire sample for households.

Figure 11: Household's Decision to Remain in the RKI-SOEP Survey (Pre-Omicron)

Figure 12: Household's Decision to Participate in the RKI-SOEP Survey (Pre-Omicron)

Figures 13 and 14 show the results for the individual's decision to remain and to take part in the study. The decision to remain appears to be only driven by a handful of variables, namely by age, citizenship, and a few select household characteristics. Similar to above, the decision to take part is strongly driven by age and other demographic variables as well as socio-economic household and individual characteristics.
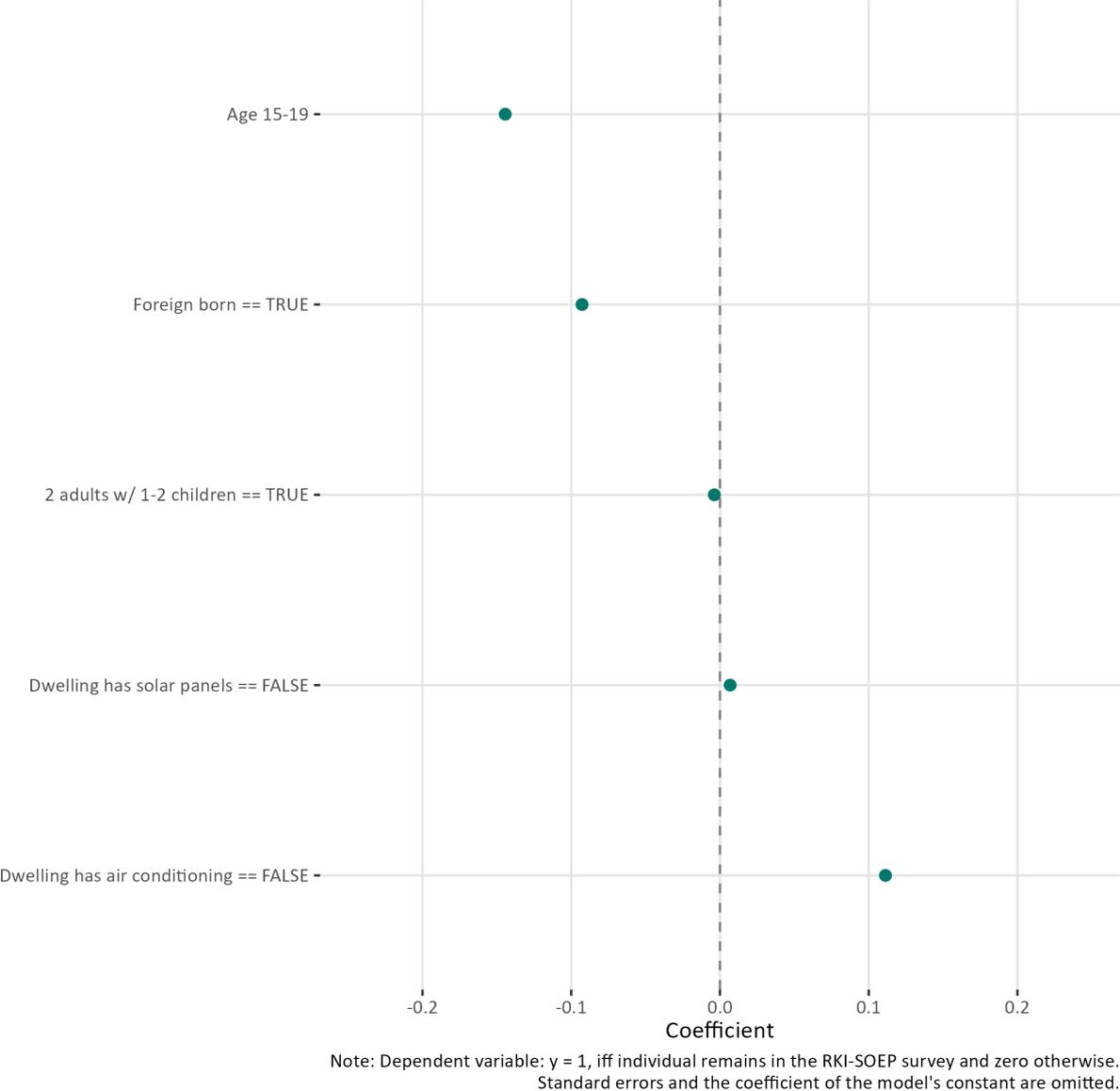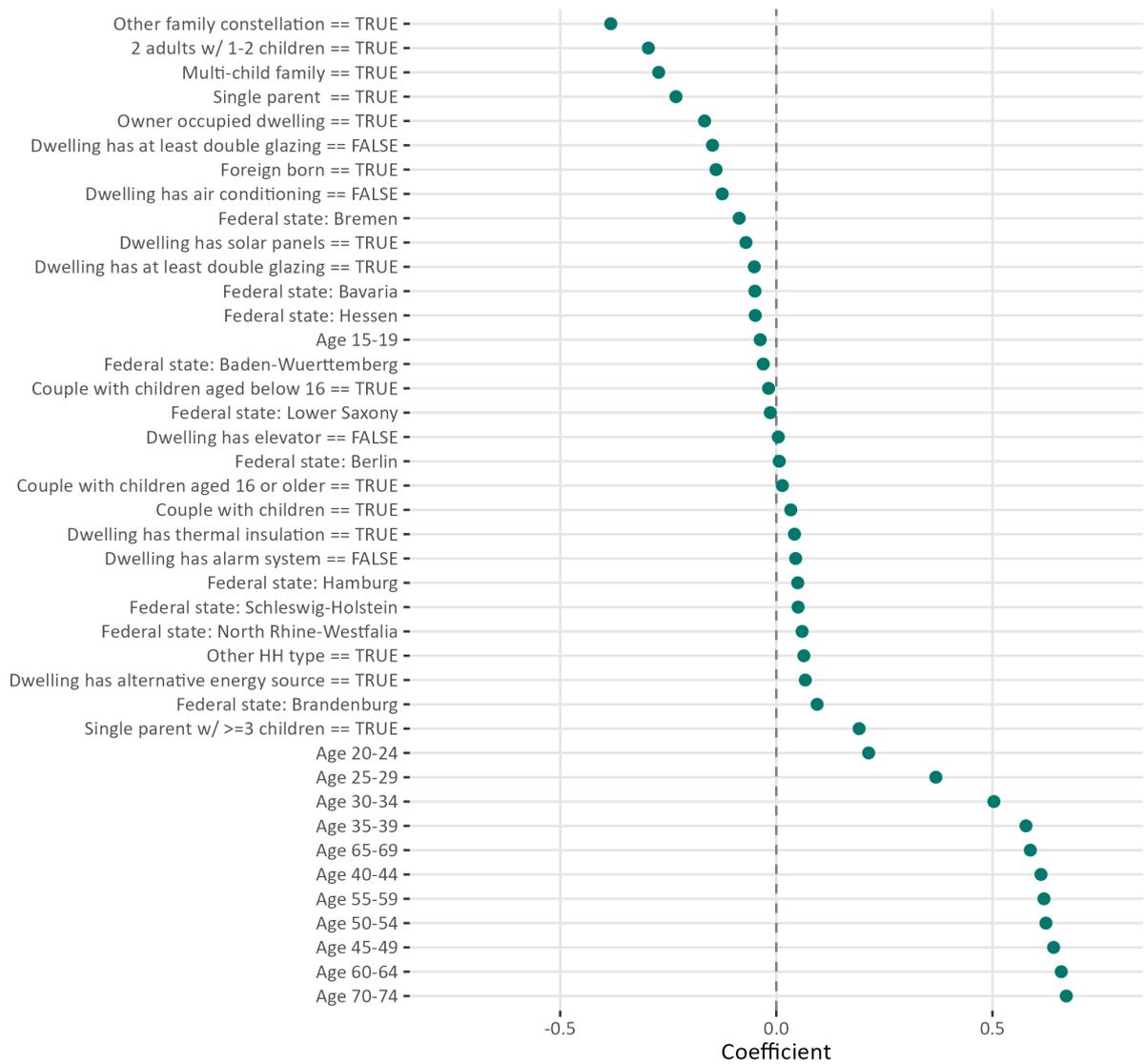

Note: Dependent variable: y = 1, iff individual remains in the RKI-SOEP survey and zero otherwise.
Standard errors and the coefficient of the model's constant are omitted.

Figure 13: Individual's Decision to Remain in the RKI-SOEP Survey (Pre-Omicron)

Note: Dependent variable: y = 1, iff individual participates in the RKI-SOEP survey and zero otherwise. Standard errors and the coefficient of the model's constant are omitted.

Figure 14: Individual's Decision to Participate in the RKI-SOEP Survey (Pre-Omicron)

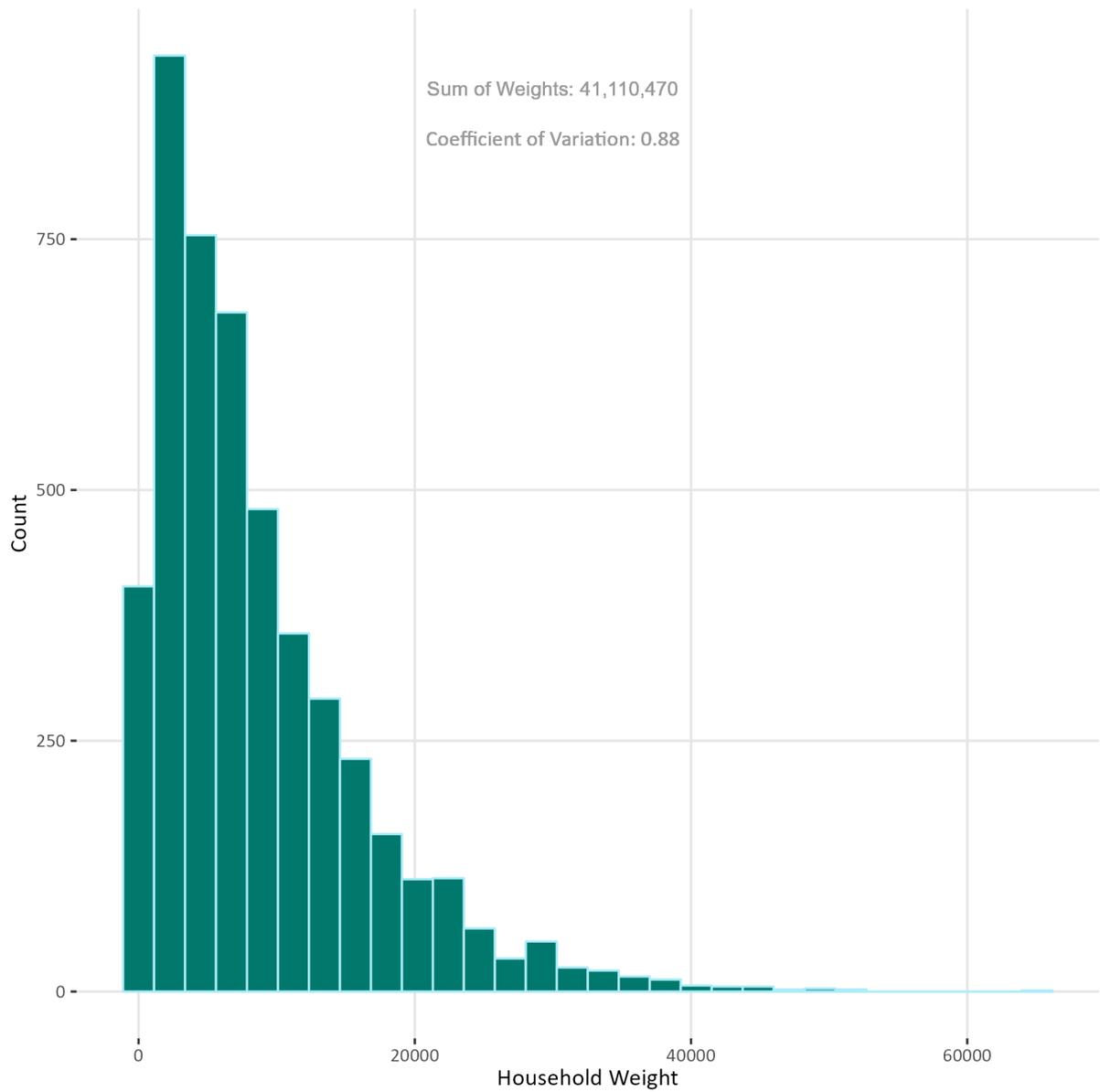Figures 15 and 16 show the resulting household and individual weights, respectively.

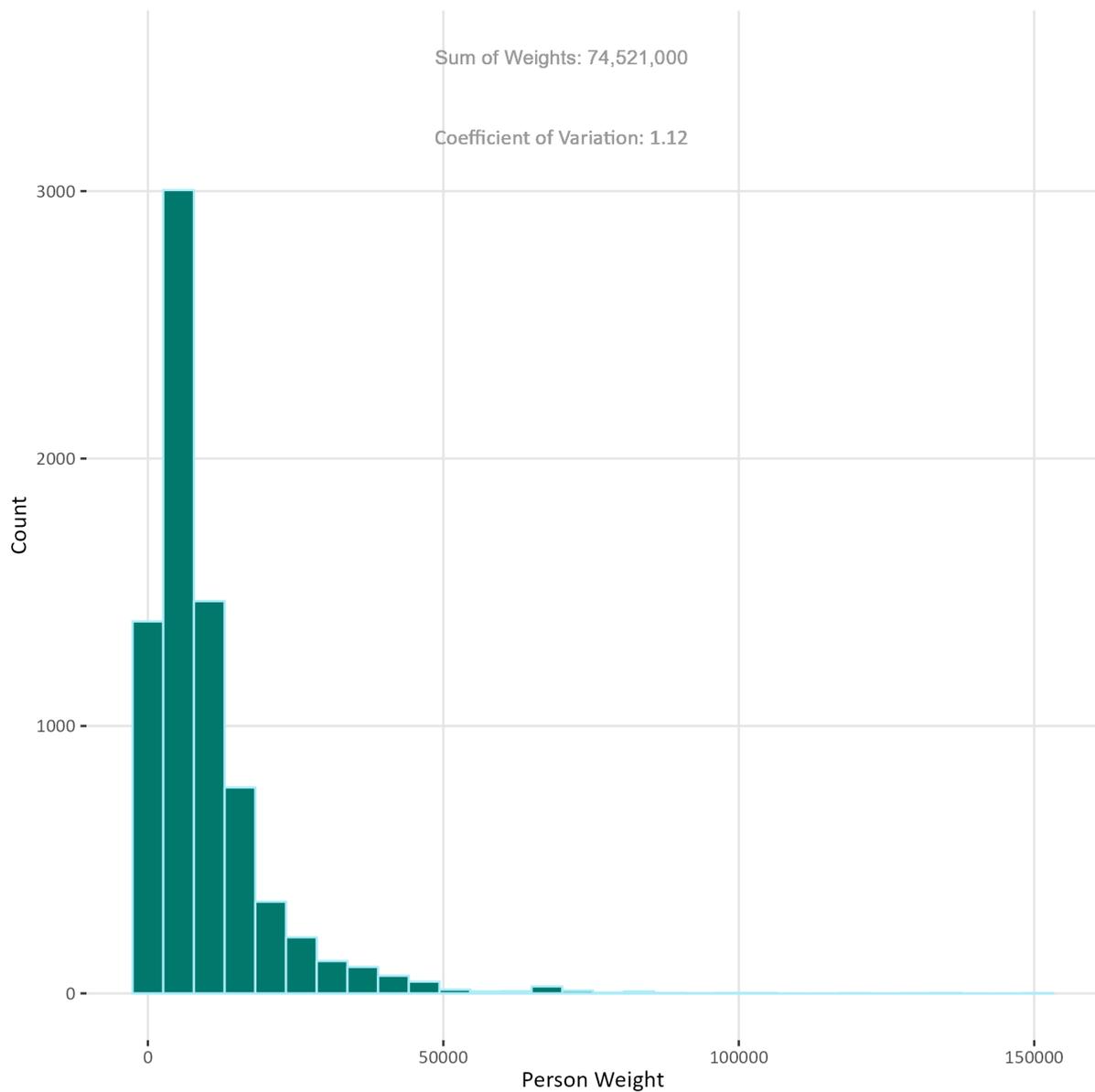Figure 15: Distribution of RKI-SOEP2 Household Weights (Pre-Omicron)

Figure 16: Distribution of RKI-SOEP2 Individual Weights (Pre-Omicron)

Figures 17 and 18 show the results for the decision to remain and to participate for households of the SOEP Core sample. In both cases, the decisions at the household level are driven by a large set of variables. The decision to remain in the study is largely driven by socio-economic and demographic factors at the household and district level. The decision to participate, is mostly driven by personal traits of the head of the household and its members as well as socio-economic factors.
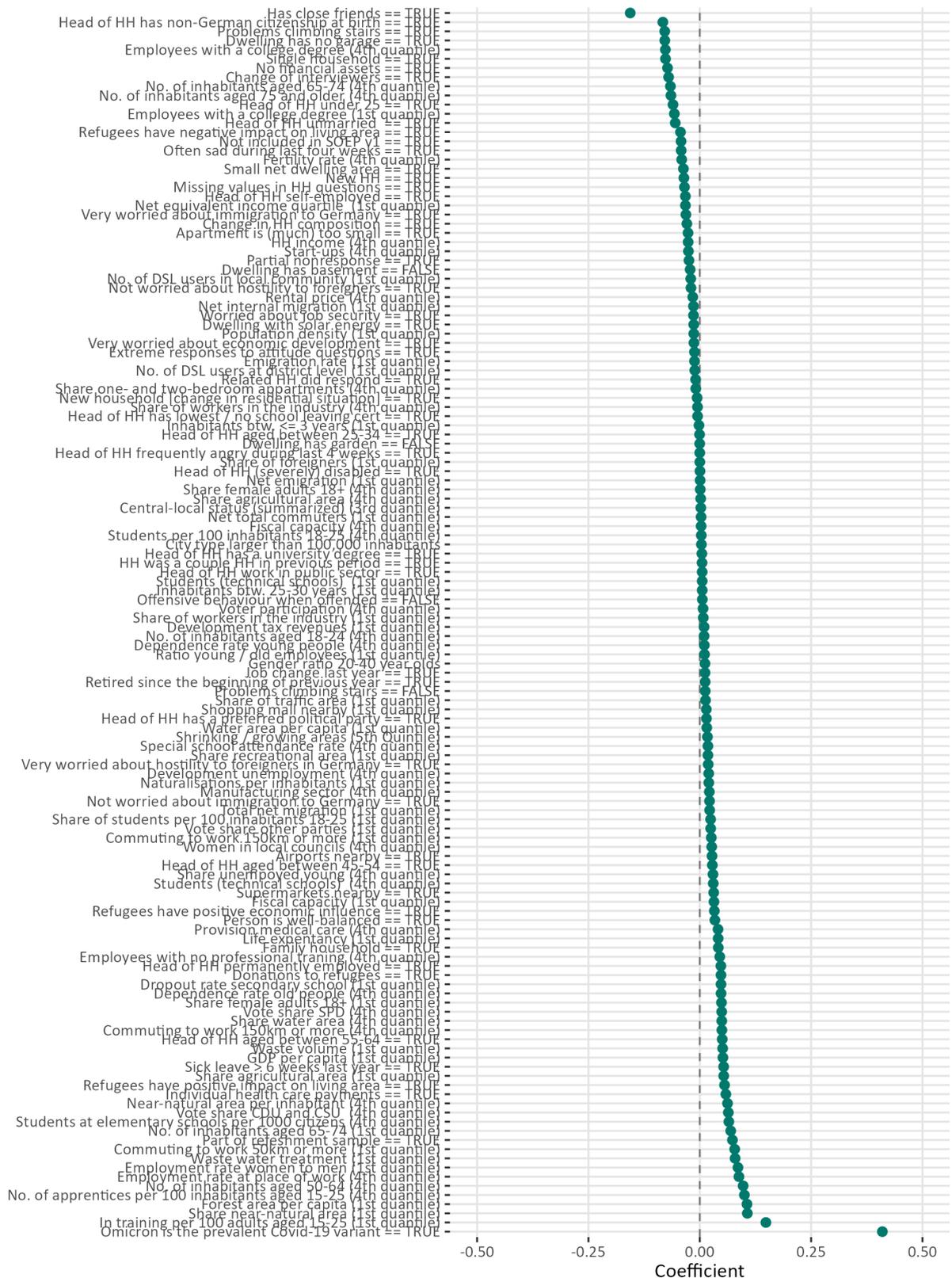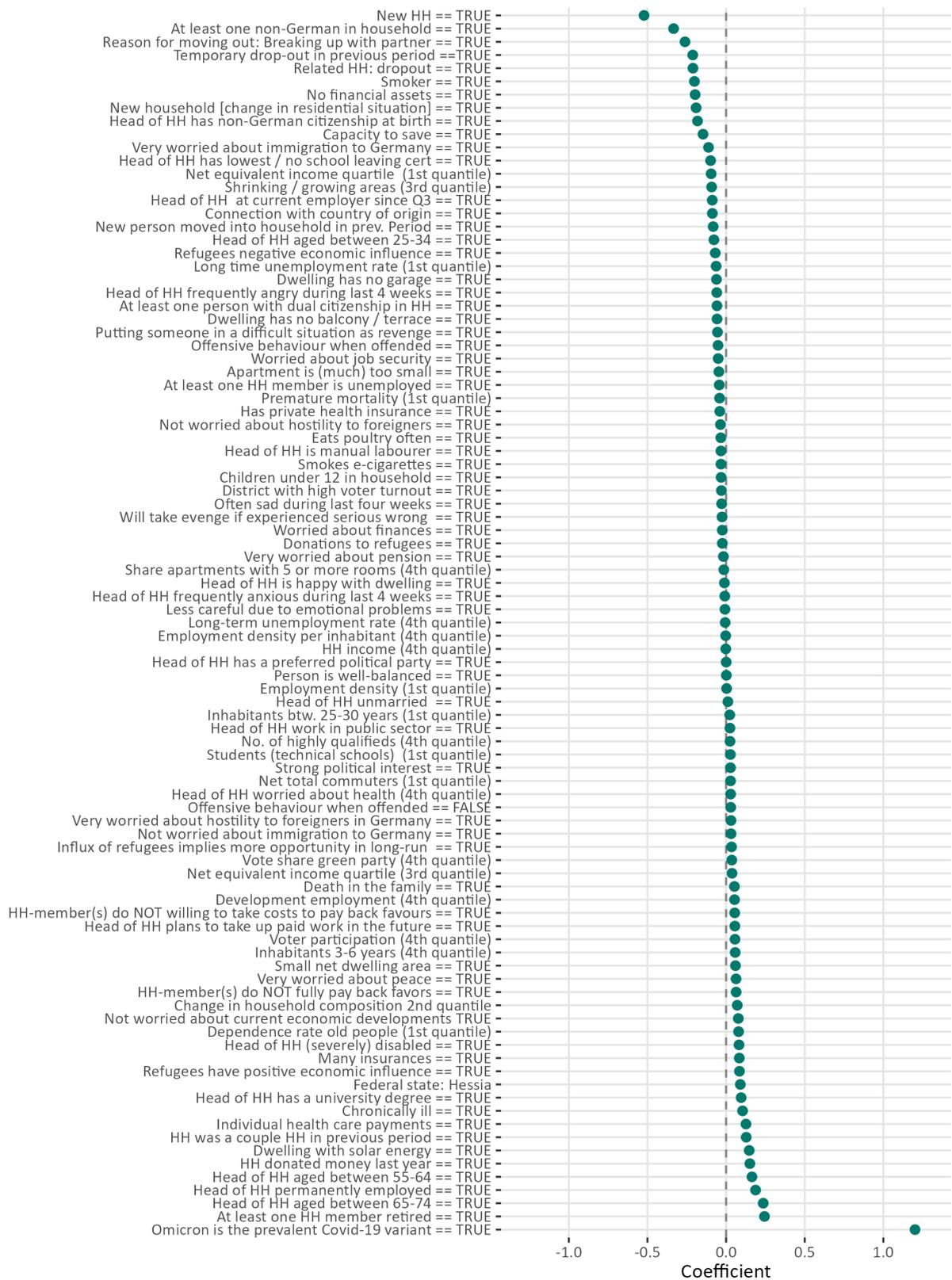
Note: Dependent variable: y = 1, iff household remains in the RKI-SOEP survey and zero otherwise.
Standard errors and the coefficient of the model's constant are omitted.

Figure 17: Household's Decision to Remain in the RKI-SOEP Survey (SOEP Core)

Note: Dependent variable: y = 1, iff household participates in the RKI-SOEP survey and zero otherwise. Standard errors and the coefficient of the model's constant are omitted.

Figure 18: Household's Decision to Participate in the RKI-SOEP Survey (SOEP Core)

Figures 19 and 20 show the individual level results for the SOEP Core sample participants. For the decision to remain, only the tranches in which the individuals were grouped appear to matter. For the actual decision to participate a large set of demographic and socio-economic variables appear to matter.
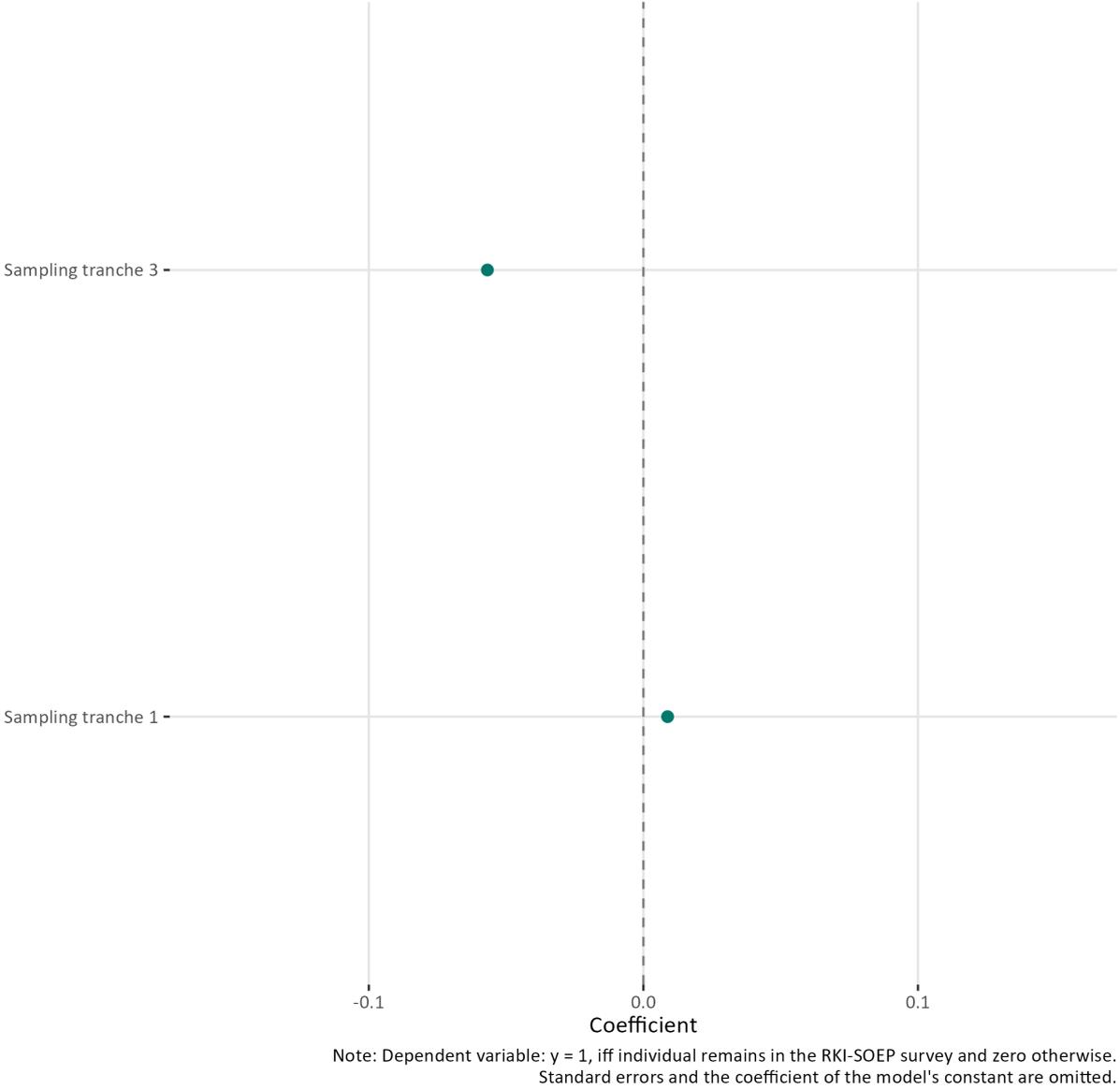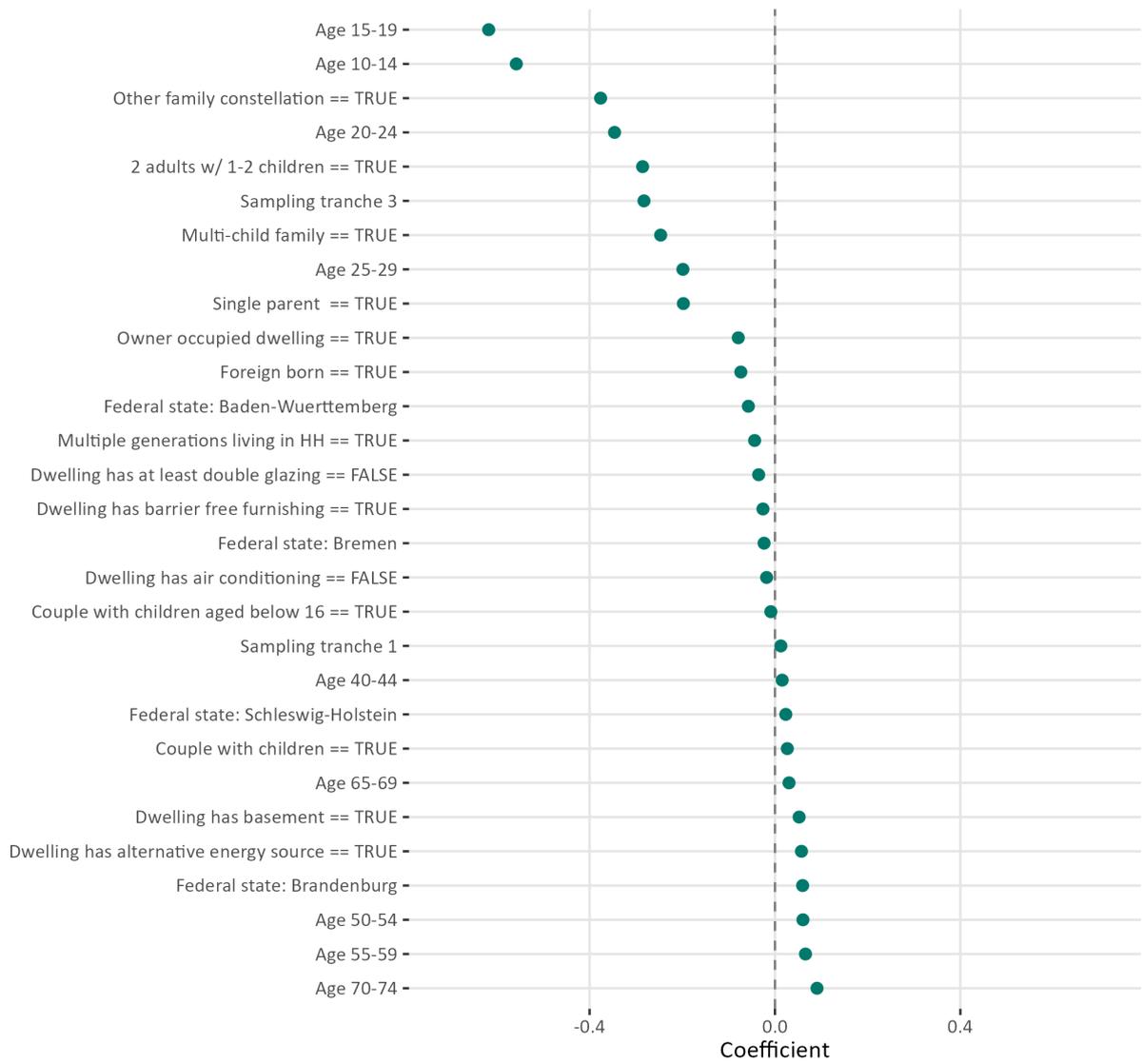

Note: Dependent variable: y = 1, iff individual remains in the RKI-SOEP survey and zero otherwise.
Standard errors and the coefficient of the model's constant are omitted.

Figure 19: Individual's Decision to Remain in the RKI-SOEP Survey (SOEP Core)

Note: Dependent variable: y = 1, iff individual participates in the RKI-SOEP survey and zero otherwise. Standard errors and the coefficient of the model's constant are omitted.

Figure 20: Individual's Decision to Participate in the RKI-SOEP Survey (SOEP Core)

Figures 21 and 22 show the distributions of the resulting household and person weights of the SOEP Core sample for reference.
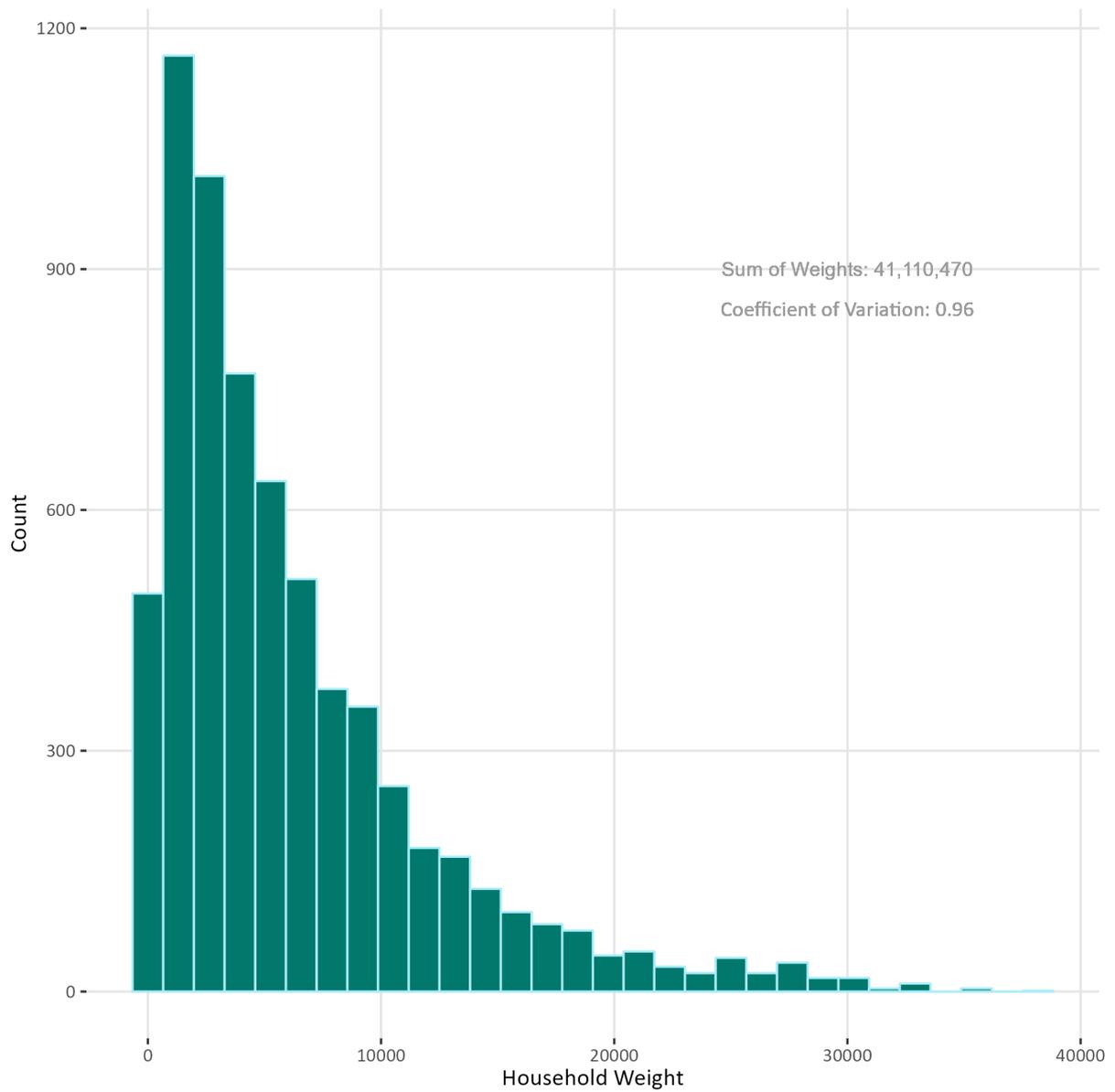
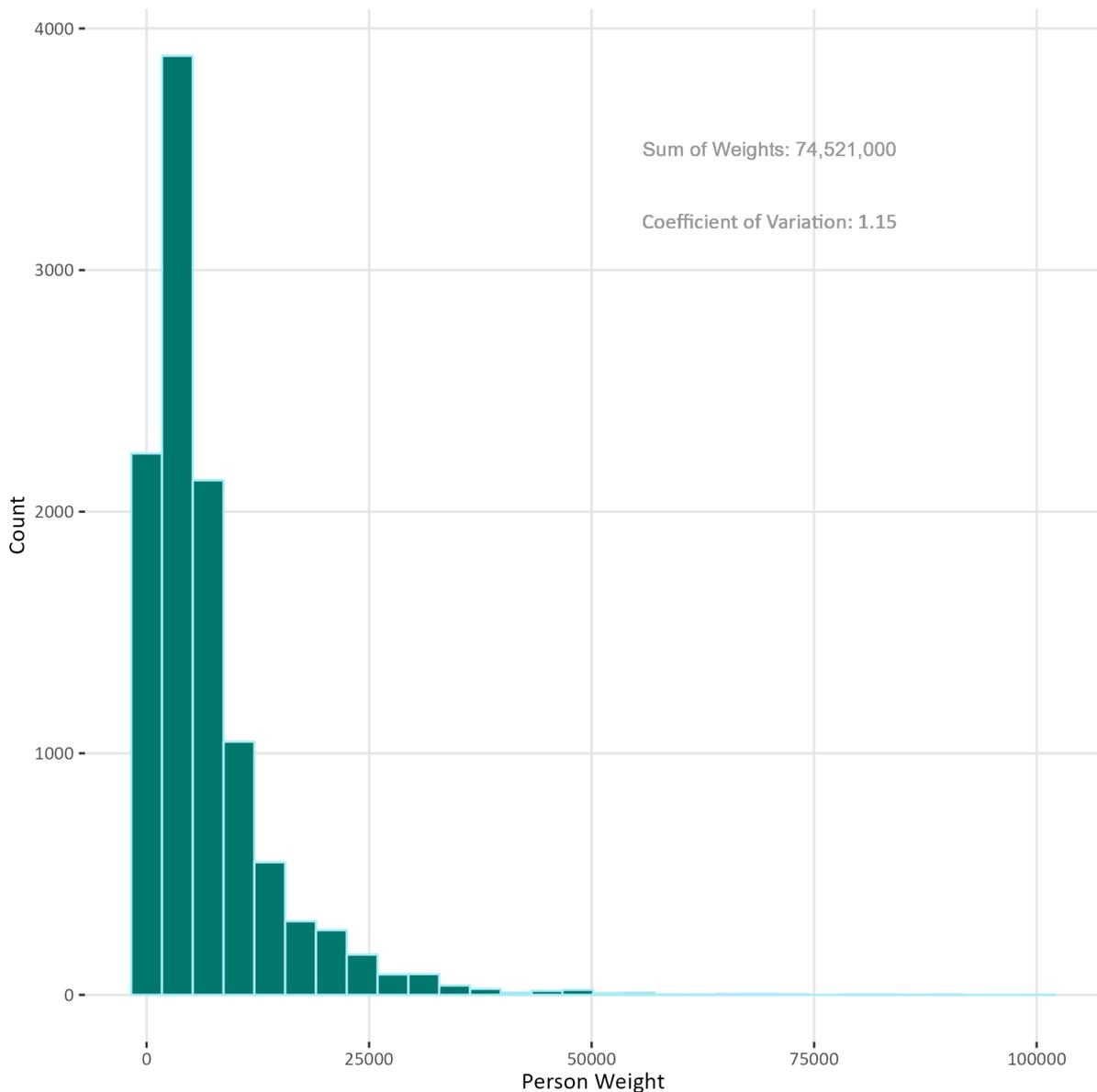Figure 21: Distribution of RKI-SOEP2 Household Weights (SOEP Core)

Figure 22: Distribution of RKI-SOEP2 Individual Weights (SOEP Core)

8 Conclusion

This paper describes the weighting methodology for the second wave of the Living in Germany - Nationwide Corona Monitoring study (RKI-SOEP2). Information from the larger, ongoing German Socio-Economic Panel (SOEP) study, from which the households were sampled, is used in order to analyze the determinants of non-contact, attrition, and refusal to participate at the household and individual levels as well as to correct the data for the results accordingly. For a sample of 6,791 households and 11,223 individuals, we find that personal attributes and socio-economic factors at the household level as well as the advent of the omicron variant are the key determinants at household level for non-response. At the individual level, demographic and socio-economic factors represent the most important predictors for non-response.

References

Chen, T., J. Kuang, R. L. Valliant, and M. R. Elliott. 2018. "Model-Assisted Calibration of Non-Probability Sample Survey Data Using Adaptive LASSO." *Survey Methodology* 44 (1): 117–27.

DeVille, J. C., C. E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88 (423): 1013–20.

Goebel, J., M. M. Grabka, M. Liebig S. Kroh, D. Richter, C. Schröder, and J. Schupp. 2019. "The German Socio-Economic Panel (SOEP)." *Journal of Economics and Statistics* 239 (2): 345–60.

Hess, D., and S. Torregroza. 2022. "Methodenbericht: Leben in Deutschland - Corona Monitoring (RKI-SOEP-2)." Infas Institut für angewandte Sozialwissenschaft.

Hoebel, J., M. A. Busch, M. M. Grabka, S. Zinn, J. Allen, Gößwald, and J. Wernitz. 2021. "Seroepidemiologische Studie Zur Bundesweiten Verbreitung von SARS-CoV in Deutschland: Studienprotokoll von Corona-Monitoring Bundesweit (SOEP-RKI)." *Journal of Health-Monitoring* 6 (S1): 2–17.

Liebig, S., H. Brücker, J. Goebel, M. M. Grabka, C. Schröder, and S. Zinn. 2021. "IAB-SOEP Migrationsstichprobe 2019." *SOEP Socio-Economic Panel Study.*

Nebelin, J., M. Petrenz, and K. Wenzig. 2019. "Wegweiser Zur Datenanalyse Der IAB-SOEP Migrationsstichproben (M1, M2) Und Der IAB-BAMF-SOEP Stichprobenvon Geflüchteten (M3-M5)." *SOEP Survey Paper 600* Series G.

Rendtel, U., R. Meister, J. Goebel, A. Gößwald, M. M. Grabka, J. Hoebel, and M. Schlaud. 2020. "Ein interdisziplinäres Studienkonzept zur Dynamik von Covid-19 auf der Basis der prospektiv erhobenen Daten der Kohorten des Sozioökonomischen Panels (SOEP)." *SOEP Papers on Multidisciplinary Panel Data Research 1094.*

Siegers, R., V. Belcheva, and T. Silbermann. 2020. "SOEP-Core V35 Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP)." *SOEP-Survey Papers 826* Series C.

Steinhauer, H. W., S. Zinn, and R. Siegers. 2021. "SOEP-Core-2020: Sampling, Nonresponse, and Weighting in Living in Germany-Corona Monitoring (RKI-SOEP)." *SOEP Survey Papers 1076* Series C.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society B* 58: 267–88.