JYVÄSKYLÄN YLIOPISTO
MATEMATIIKAN JA TILASTO-
TIETEEN LAITOS

Master's thesis

A Bayesian two-part model for improving social assistance estimation of the SISU microsimulation model

Annakaisa Ritala

June 20, 2024

Author
Annakaisa Ritala

Title
A Bayesian two-part model for improving social assistance estimation of the SISU microsimulation model

Degree program
Master's Degree Programme in Statistics and Data Science

Date
June 20, 2024

Page count
65 pages, including 8 appendix pages

Tiivistelmä

SISU–mikrosimulointimalli on keskeinen väline sosiaaliturvaetuuksien ja tuloverotuksen lainsäädännön valmistelussa Suomessa. Mikrosimulointi tarkoittaa, että kiinnostuksen kohteena on laskea ennusteita havaintoyksiköille, jotka SISU–mallin rekisteriaineistossa ovat yksilö ja asuntokunta. SISU–mallilla voidaan simuloida lakimuutosten yhteisvaikutuksia sosiaaliturvaetuuksiin, mukaan lukien perustoimeentulotukeen. Toimeentulotuki on kuukaudelta myönnettävä viimesijainen etuus, jonka tavoitteena on varmistaa perheen ihmisarvoisen elämän kannalta välttämätön toimeentulo. SISU–mallin perustoimeentulotuen ennusteissa on kuitenkin epätarkkuutta. Tämän pro gradu -tutkielman tavoitteena oli tutkia voiko SISU–mikrosimulointimallin vuosittaisia toimeentulotuen ennusteita tarkentaa Bayes-tilastotieteen menetelmien avulla.

SISU–mallin perustoimeentulotuen ennusteiden epätarkkuus syntyy mahdollisesti useista lähteistä. Esimerkiksi SISU-mallin hyödyntämän rekisteriaineiston muuttujat on kirjattu vuositasolla, vaikka toimeentulotuki myönnetään kuukaudelta. Lisäksi rekisteriaineistosta puuttuu toimeentulotuen suuruuden määritykseen käytettäviä muuttujia, kuten terveysmenot ja varallisuus. Myös SISU–mallin oletus, että kaikki toimeentulotukeen oikeutetut asuntokunnat hakisivat etuutta ei vastaa todellisuutta.

Tutkielmassa kehitettiin kaksiosainen malli, joka koostui kahdesta yleistetystä lineaarisesta mallista. Kaksiosaisen mallin rakenteen avulla voitiin yhdistää aineiston generoivan prosessin kaksi osaa. Ensimmäinen prosessi mallinsi todennäköisyyttä saada toimeentulotukea ja toinen prosessi vuosittaista toimeentulotuen määrää, ehdolla, että asuntokunta saa toimeentulotukea. Ennusteiden simulointiin käytettiin Bayes-menetelmiä.

Kehitettyjä kaksiosaisia malleja ja SISU–mallia vertailtiin luokitteluvirhettä ja ennustetarkkuutta mittaavien tunnuslukujen suhteen. Löydösten perusteella

kehitettyjen mallien toimeentulotuen saannin ja määrän ennusteet olivat tarkempia kuin SISU–mallin, riippuen mitä jakaumaa toimeentulotuen määrän oletettiin noudattavan ja minkälaista luokitterurajaa käytettiin toimeentulotuen saannin ennustamiseen.

Tutkielma osoittaa, kuinka tilastollisen mallinnuksen avulla voidaan huomioida käyttäytymiseen ja aineiston puutteisiin liittyvää epävarmuutta toimeentulotuen ennustamisessa, ja kuinka mallinnuksen avulla on mahdollista saada tarkempia toimeentulotuen ennusteita kuin SISU–mallilla. Vaikka kehitettyihin malleihin ei sisällytetty lakimuutosten vaikutusten simuloinnin mahdollistavaa mekanismia, työ syventää käsitystä SISU–mallin kehityskohteista ja mikrosimuloinnin laajentamisesta tilastollisen mallinnuksen avulla.

Contents

Introduction

The SISU microsimulation model is an essential tool for assessing the annual costs of social benefit programs and the impact of legislation reforms in Finland. The SISU model is developed and maintained by Statistics Finland and a description of the model can be found in the SISU microsimulation manual (Statistics Finland, 2018). Microsimulation refers to a simulation method where the units of analysis are unique entities, and in the SISU model, the units of analysis are individuals and house-dwelling units. In brief, the SISU model is a program which contains the Finnish social security legislation. The model may be used to estimate the social benefits of interest received by an individual or a house-dwelling unit. It is currently used to provide estimates for tax and social security legislation preparations and assess the impact of legislation reforms on the annual social security budget and income distribution.

Basic social assistance is one of the social benefits which may be simulated with the SISU model. The rules governing social assistance are determined through political decision-making, dictating both eligibility criteria and the extent of the benefit. According to the Social Assistance Act (SAA 1412/1997) social assistance is a last-resort financial aid and the purpose of the benefit is to ensure a family's minimum income and advance income independence (SAA 1.1 §). The benefit is aimed to be granted temporarily for a period of 1–2 months. In Finland, social assistance consists of three parts: basic, supplementary and preventative social assistance (SAA 6–8 §). In simple terms, the amount of social assistance granted to a family is calculated as a difference between the income and expenditure of the family (SAA 6.1 §).

It is in the interest of governmental agencies and research institutions to be able to predict and evaluate the impact of legislative reforms on the number of social assistance recipients and the amount received by families annually. For example, the National Pension Index Act 1064/2010 4 a § defines a directive to evaluate the adequacy of basic social security every four years. Basic social security refers to the minimum level of benefits a person receives when they are not working, and social assistance is an important tool to ensure the adequacy of basic social security (Tervola et al., 2023). Further, social assistance is funded by the government and municipalities (SAA 27 f §). A review on social assistance expenditure in 2019 by Tanhua and Kiuru (2020) showed that the total expenditure on social assitance was 780.5 million euros and it was granted to 9.5% of all Finnish households. Basic social assistance formed 92% of the total expenditure, whereas supplementary and preventative social assistance formed around 4% and 3% of the total expenditure, respectively. In brief, these impact evaluations are of interest because the social assistance legislation in part defines what is considered as adequate expenditure to ensure a basic livelihood, and on the other hand, the provision of social assistance

imposes costs on the government and the municipalities.

However, at present Statistics Finland (2018) reports that the SISU model produces somewhat unreliable estimates of social assistance. Some reported reasons for the inaccuracies are the quality of the register data and the units of analysis. For example, the register data set uses house-dwelling units as units of analysis which conflicts with the definition of a family in the social assistance legislation, the variables are recorded on an annual level whereas social assistance is granted on a monthly basis, and the data set does not contain all variables relevant for the social assistance determination. Moreover, the SISU model calculates the social benefits deterministically and does not estimate uncertainty arising from the scarce information or uncertainty resulting from the take-up behaviour. *Take-up behaviour* refers to decision-making whether to apply for a social benefit (Kuivalainen, 2007).

Some systematic error in the basic social assistance estimates could be addressed by modelling the data-generating process using a two-part model. In essence, the two-part model (Liu et al., 2019; Neelon et al., 2016) combines the two parts of the data-generating process: one that determines the probability of receiving social assistance and another that determines the annual amount of social assistance. Further, by implementing these two processes using generalized linear modelling (Fahrmeir et al., 1994), the uncertainty over the receipt status and the distribution of annual social assistance may be estimated from the data. Therefore, this master's thesis aims to develop a two-part model to improve upon house-dwelling unit-level annual social assistance estimates by using Bayesian statistical methods. The Bayesian framework was selected to incorporate parameter uncertainty into the model predictions and measure this uncertainty as probabilities.

The main significance of the following study is to demonstrate how the two-part mixture modelling approach describes the data-generating process of social benefit programs including social assistance. The advantage of the framework is that it may consider uncertainty arising from scarce information and the take-up behaviour of potential recipients. As a result, there is potential to improve the predictive accuracy of annual social assistance given that the legislation is fixed. Further, the model adds to research investigating how statistical modelling can supplement microsimulation. Finally, the presented modelling framework could pave the way for the incorporation of more advanced statistical methods, such as causal modelling of social benefits or modelling the development of social benefits over time.

First, in Section 1, the social assistance legislation is summarised and in Section 2, the SISU microsimulation model, register data set, current estimation method of social assistance and current predictive accuracy of the SISU model are reviewed.

In Section 3, characteristics associated with social assistance receipt and the distribution of social assistance according to family characteristics are described. In Section 4, the theory of two two-part models is introduced, which lays a foundation for modelling the data-generating process of social assistance. Further, in Section 5, the model estimation and selection are explained, including the external validation procedure, selection criteria and sensitivity analysis. In Section 6, the results of the developed two-part models are presented. Lastly, in Section 7, the model results are interpreted, and their implications are discussed.

Artificial intelligence-based applications, Grammarly and ChatGPT, were used in this study to improve grammar, diction, and clarity.

1 Social assistance legislation

The following section summarises the Social Assistance Act in 2024 (SAA 1412/1997). Social assistance is a last-resort financial aid. The purpose of the benefit is to ensure a family's minimum income and advance income independence (SAA 1.1 §). The benefit is aimed to be temporary: it is intended to be granted on a monthly basis, one to two months at a time (SAA 13.1 §). Social assistance is granted to the applicant's family. A family may consist of parents living in a common household, a parent's minor child, married couples or couples in circumstances similar to marriage who live in a common household (SAA 3.1 §). Social assistance is considered to be shared equally among each family member (SAA 3.2 §).

In Finland, social assistance consists of three parts: basic, preventative and supplementary social assistance. The basic part covers basic needs, mainly housing costs and health care costs (SAA 7 §), and it is granted by the Social Insurance Institute of Finland (fin. Kela). Supplementary social assistance may be granted to cover special needs or situations, such as long-term illness (SAA 7 c §). Lastly, preventative social assistance may be granted to prevent long-term reliance on social assistance (SAA 8 §). Both supplementary and preventative benefits are discretionary, and these benefits are granted by wellbeing service counties (SAA 14 c §). A family may apply for supplementary and preventative social assistance after applying for basic social assistance regardless of whether the application for basic social assistance was successful (SAA 14.2 §).

Basic social assistance is means-tested: the Social Assistance Act determines the criteria that members of the applicant's family must meet to receive full social assistance. First, each applicant who is 17–64 years of age must sign up as unemployed, unless they are working, studying full-time or are unable to work (SAA 2 a §). The failure to meet set criteria leads to reductions in the basic amount of the aid (SAA 10 §). Second, the applicant must apply for all the other social benefits they are eligible for (SAA 6.3 §). In simple terms, the amount of

social assistance granted is a difference between the costs and income of the family (SAA 6.1 §). However, some income sources are exempt from the calculation, and the Social Assistance Act details the intended expenses of a family. Some details about how the amount of basic social assistance is determined are presented next.

The basic social assistance legislation determines the expected level of expenses which the benefit covers based on the family structure. These expenses are considered as the costs in the calculation of the received final basic social assistance amount. The expected level of expenses consists of a basic amount, living costs and basic health care expenses (SAA 7.1 §). The *basic amount* is a predetermined amount that represents the intended monthly expenditure on basic needs for a given family structure (SAA 9 §). The basic amount is the intended expenditure on costs of food, clothing, hygiene products, phone bills and recreational expenses, such as hobbies. The baseline of the basic amount is determined for a lone-dweller, and the other family compositions' basic amount is determined by scaling this amount with a corresponding weight (SAA 9 §). The basic amount was 497.29 euros per month for a lone-dweller in 2019 after index increases (SAA 9 §). In addition to the basic amount, basic social assistance covers living costs such as rent (SAA 7 a §) and health care costs (SAA 7 b §) to an appropriate extent. The basic amount is adjusted annually based on the national pension index (SAA 9 a §).

The basic social assistance legislation defines which income sources and expenditures are considered in the social assistance calculation. Both the applicant's and their family members' incomes are taken into consideration. There is an income disregard of 150 euros from an earner's wage (SAA 11.3 §). In addition, some further income is exempt, for example, an under-18-year-old child's wages (SAA 11.1,2 §) and income covering travel and expenses related to work (SAA 11.1,3 §).

2 The SISU microsimulation model

In the following section, the descriptions concerning the SISU model are from the SISU microsimulation manual (Statistics Finland, 2018). The SISU model is an SAS program, which is used to estimate social benefits received by individuals and house-dwelling units on an annual level using rules defined in the social security legislation. As its input, the program takes a register data set, the legislation year and legislation-specific parameters and it generates social benefit estimates using rules coded into its' subprograms.

The SISU model is deterministic and static. Static means the SISU model estimates immediate effects and does not take into consideration possible temporal dependency of social assistance receipt in time such as a previous year or a previous

month. Further, the SISU model is deterministic, meaning it does not consider the impact of take-up behaviour or possible behavioural changes elicited by legislative reforms. Additionally, the SISU model does not measure uncertainty introduced, if some variables needed to calculate a social benefit of interest are missing. However, the future developments of the SISU model aim to take into account possible behavioural and temporal impacts of reforms.

The SISU model has been developed to aid in the preparation of tax and social benefit legislation, for example, to estimate the budget and income distribution effects of the legislative reforms (Tervola et al., 2023; Kotamäki et al., 2017), and to monitor the fulfillment of intended effects of previously enacted legislation (Mesiäislehto et al., 2022). For example in a study by Kotamäki et al. (2017) the SISU model was used to assess the impact of a potential 100-day cut to the duration of the earnings-related unemployment benefit in 2017 on employment, public sector finances and the income distribution. The SISU model was used to evaluate the impact of the reform by calculating earnings-related unemployment benefit totals without the reform and the counterfactual impact of the reform. The impact was assessed by investigating changes in the number of affected individuals and changes in disposable income in groups defined by, for example, income decile, sex, and family type. Another example of estimation of income distribution effects is that the adequacy of basic social security is assessed every four years, and the SISU model has been used to estimate the effects of the relevant social security and taxation reforms on basic social security (Tervola et al., 2023). The effects were measured, for example, by calculating annual poverty indices in groups defined by age and sex while taking into consideration the general price level (Tervola et al., 2023, see Table 8.1). In general terms, the SISU model may be used to simulate alternative scenarios and compare outcomes of interest between these scenarios while keeping other variables of interest constant.

2.1 Register data set

At the start of the study, the latest available register data set was the 2021 register data set. Therefore, the 2019 register data set was chosen for model development because the years 2020 and 2021 were considered atypical regarding social security expenditure. The review by Isotalo et al. (2022) reported that the COVID-19 pandemic caused a short decline in economic activity during the spring and summer of 2020. Further, the review showed that there was an increased number of furloughs and a moderate increase in the unemployment rate, and the employment rate did not return to the 2019 level until September 2021. Especially at the beginning of the pandemic, the number of people receiving unemployment benefits was 1.75 times higher in May 2020 than in May 2019 (Isotalo et al., 2022). Therefore, it was thought that 2019 would be the most representative of typical social security

expenditure in the future.

Moreover, note that the current social assistance legislation in 2024 largely conforms to the social assistance legislation that was in effect in 2019. A notable change in 2023 was that the execution of preventative and supplementary social assistance was allocated to wellbeing service counties (1023/2022), whereas in 2019, the execution of these benefits was conducted by municipalities.

The following overview summarises the descriptions and sampling of the register data set in the SISU microsimulation manual (Statistics Finland, 2018). The descriptions concerning the 2019 register data set have been generated by the author.

The register data set used by the SISU model is collated for purposes of microsimulation. With approximately a one-year lag, the SISU model is updated with the newest available register data set and the corresponding legislation. The 2019 register data set contains observations from 828958 individuals consisting of 432962 house-dwelling units. The sample contains 15% of the population living in Finland at the end of the reporting year, and the data set's units are an individual and a house-dwelling unit. A *house–dwelling unit* consists of individuals living at the same address. Homeless persons, persons living in supported housing or persons who do not have an address are included in the register data set with a lone-dweller status. Further, the sampling frame contained persons whose domicile was Finland on the last day of the year according to the Population Information System. The sampling is conducted as a systematic sampling of the sampling frame, where sampling units have been ordered according to disposable money income (Statistics Finland, 2018). This allows for an improved representation of high-income families in the data set.

The register data set is collated from administrative data sets and registers, such as the person and property database by Statistics Finland and the social assistance register data set by the Finnish Institute for Health and Welfare. Moreover, some missing information has been imputed, such as some housing costs. For example, missing rent information was imputed using a matching method. In the matching method, the distances between the observations were calculated using a distance measure, which measured the differences between the observations based on the area of the apartment, geographical location, and housing tenure. The specific formulation of the distance measure was not presented. The imputed rent value was selected from the nearest house-dwelling unit based on the distance measure.

The 2019 register data set has 214 variables which describe the individual or the house-dwelling unit (e.g. age, number of persons living in the house-dwelling unit), their potential social benefits and other income, such as salary. In case the house-dwelling unit is selected as the unit of analysis, the variables describing an

individual such as age or education level, are reported from the reference person. The *reference person* in the family is the member, whose annual net income is the highest (Statistics Finland, 2018). In case the net income does not define the reference person, the eldest family member is the reference person. The data is recorded on an annual level, meaning units' values are recorded as annual totals and point estimates in a single month, such as the annual total of housing allowance received or the main activity of an individual on the last day of the year.

In the register data set, 90.9% of the house-dwelling units are not social assistance recipients, and respectively 9.1% receive some form of social assistance. In general, the distribution of empirical annual social assistance forms a continuous, positive and right-skewed distribution (Figure 1). Among social assistance recipient house-dwelling units, the median annual social assistance was 1916 euros and the mean was 2823 euros. The first quartile was 742 euros and the third quartile was 4016 euros. This supports previous findings (Tanhua and Kiuru, 2020; Jauhiainen and Korpela, 2019) that most of the social assistance recipients receive rather small amounts of the benefit annually.

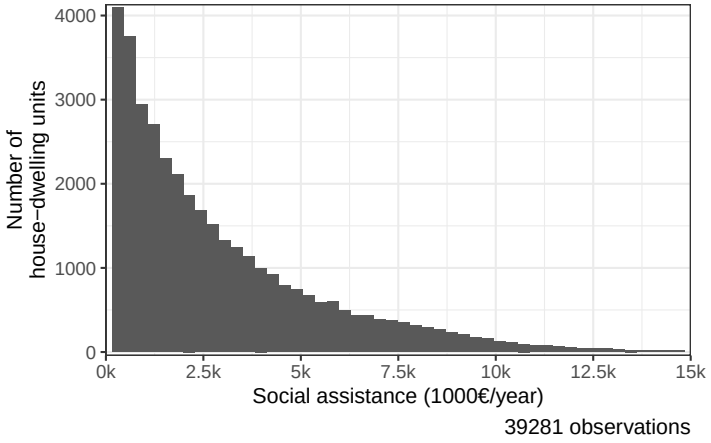

Figure 1: The empirical distribution of the annual social assistance in the 2019 register data set. The number of observations indicates the number of visualized observations. Only positive observations are displayed due to the large proportion (90.9%) of zero observations. The right tail of the distribution has been truncated at 15000€, resulting in the exclusion of 146 positive observations.

2.2 Estimation method of basic social assistance

The SISU model estimates the basic social assistance via a social assistance submodel. The basic social assistance estimation is not restricted to those who are recipients in the 2019 register data set, but the amount of benefit is estimated for all potential recipients (Statistics Finland, 2018). The estimation of the social

assistance has been adapted to be compatible with the annual level of the data and missing variables. The following explanation of the estimation method summarises the so-called "full model" subprogram of the 2019 SISU model version.

Let i denote the index of an individual house-dwelling unit, where $i = 1, \ldots, n_{\text{reg}}$ and n_{reg} is the number of observations in the register data set. First, the variables needed for social assistance calculation for the house-dwelling units are selected from the register data set, for example, the house-dwelling unit structure and annual salary. Then the income received as social aid, such as the housing allowance or unemployment benefits are estimated using their respective submodels. Monthly averages of the income variables are calculated to align with the monthly level of the social assistance legislation. Then a monthly estimate of social assistance $\hat{\mu}_{k,i}$, where k is the index of a month, is estimated according to the social assistance legislation. Third, annual average of the basic social assistance $\hat{\mu}_i = \frac{1}{12} \sum_{k=1}^{12} \hat{\mu}_{k,i}$ is calculated. Finally, an annual basic social assistance estimate $\hat{y}_{\text{sisu},i}$ is calculated such that

$$\hat{y}_{\text{sisu},i} = 12 \cdot \hat{\mu}_i \cdot r_i$$

where the constant r_i scales the amount of social assistance depending on the number of days in the military or civil service.

2.3 Estimation accuracy of the SISU model

2.3.1 Classification accuracy of basic social assistance receipt

The classification accuracy of recipient status refers here to the rate at which a model makes a correct classification of whether a house-dwelling unit receives basic social assistance. The classification accuracy was assessed through misclassification rate, false positive rate and false negative rate. The *misclassification rate* was calculated as the proportion of all observations which had an incorrectly predicted social assistance receipt status. The *false negative rate* was calculated as the proportion of social assistance recipients who were incorrectly predicted as non-recipients. The *false positive rate* was calculated as the proportion of non-recipients who were incorrectly predicted as social assistance recipients.

First, in the observed data, the basic social assistance recipient was defined as a house-dwelling unit whose amount of basic social assistance was 1 euro or higher. Otherwise, the house-dwelling unit was considered as a non-recipient. In the 2019 register data set, the false negative rate of the SISU model was 50.47% and the false positive rate was 4.48%. This means the SISU model identifies around half of the true social assistance recipients correctly and that only a small proportion of the non-recipients are classified incorrectly. These metrics suggest that the SISU model performs rather well in avoiding false alarms but has difficulties in discriminating the true social assistance recipients from the non-recipients. However, these

metrics reflect the imbalance of non-recipients and recipients in the data set, because it is a more difficult classification task to identify a social assistance recipient than it is to identify a non-recipient.

At present, the SISU model predicts a social assistance recipient status of a house-dwelling unit correctly 8.72% of the time. While this appears to be a rather good performance, the misclassification rate does not consider the distribution of social assistance recipients and non-recipients in the data set. In effect, because the incidence of social assistance receipt is around 9.10% in the 2019 register data set, a dummy classifier which predicts none of the observations are social assistance recipients would have a misclassification rate of 9.10%. This means that because the distribution of social assistance recipients and non-recipients is unbalanced, the misclassification rate should be interpreted in relation to a dummy classifier. In comparison to the dummy classifier, the classification accuracy of the SISU model in general is rather modest. Altogether, the classification performance of the model is best evaluated in terms of the false positive rate and the false negative rate, because there is a high imbalance in the proportion of the social assistance recipients and the non-recipients in the data set.

2.3.2 Predictive accuracy of basic social assistance

The predictive accuracy of annual basic social assistance refers here to the residual between the observed annual social assistance received and the annual basic social assistance predicted by the SISU model. The residuals are presented in Figure 2 in groups defined by the classification outcome, that is true positive, false negative, and false positive. True positive observations refer to observations which were observed to receive basic social assistance and were predicted to receive basic social assistance. False positive observations refer to observations which were not observed to receive basic social assistance but were predicted to receive social assistance. Respectively, false negative observations refer to observations which were observed to receive social assistance but were predicted to not receive social assistance.

The residuals in Figure 2 highlight the variability in the error regarding false negatives and false positives. First, the residuals in the true positive group in Figure 2 appear symmetric around zero, which suggests the SISU model over- and underestimates the annual social assistance at an even rate for the true positive observations. Second, the distribution of residuals for the false positive and the false negative observations describes how the classification error is connected to the residuals. The high density of small residuals for the false negative observations indicates that the SISU model has difficulty finding recipients receiving small amounts of social assistance annually. Because of the annual level of the register data set, this is an expected result, as further elaborated in Section 2.4.

On the other hand, there is a second peak in the residuals among the false positive observations near −6000 euros. The systematic pattern of the large false positives is possibly in part due to a systematic error, which is related to specific vulnerable groups who do not typically receive social assistance. Upon closer inspection of the register data set, this negative peak consists of house-dwelling units lacking any income, have unknown housing tenure and whose main activity is unknown or belong to an inactive demographic. This would suggest that the SISU model may have difficulties capturing marginal groups who do not rely on social assistance.

Altogether, it appears that the prediction error for the false negatives might largely result from the annual level of the dataset whereas the prediction error for the false positives might result from systematic error. The possible sources of error in the SISU model basic social assistance estimates are further elaborated next.

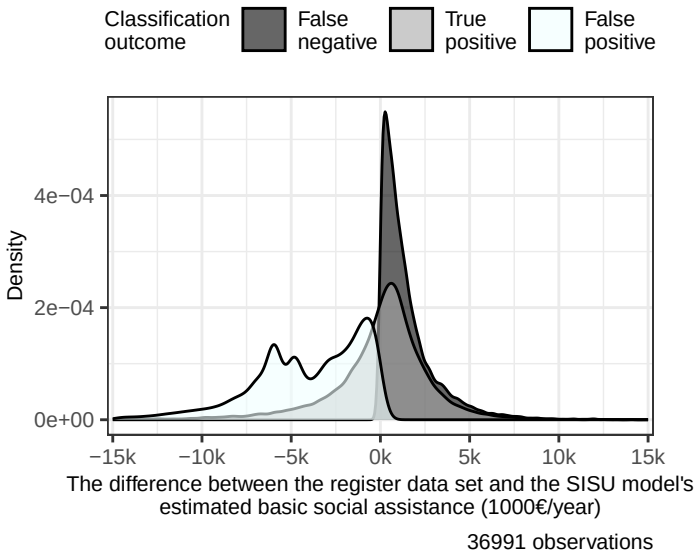

Figure 2: The difference between empirical annual basic social assistance and the SISU model estimated annual basic social assistance in the 2019 register data set. The differences were calculated by subtracting the SISU model estimates from the observed basic social assistance. The coloured regions indicate the classification outcome of each observation, that is true positive, false positive or false negative. The number of observations indicates the number of visualized observations. True negatives are removed because these observations all receive a zero value and form a large proportion (86.3%) of the observations. The left and right tails of the distribution have been truncated at −15000€ and 15000€, respectively. Truncation resulted in the exclusion of 390 observations.

2.4 Sources of bias in the SISU basic social assistance estimates

In the following section, the descriptions of the social assistance submodel are from the SISU microsimulation manual (Statistics Finland, 2018). The SISU model's social assistance estimates are biased due to the annual level of the data, missing data, the definition of units of analysis, non-take-up behaviour and due to the hierarchical organisation of the SISU model. Details of the sources and their effect on the SISU model estimates are presented next.

The register data does not include information on all income sources or family expenditures. First, there is no information on monetary benefits received from friends and family or the wealth of the family, both of which are considered when social assistance is granted. Second, there is no information on family health care expenditure or other brief increases in the expenditure of a family, such as replacing a broken household appliance. Finally, some uncertainty arises from the imputation of the housing costs. In summary, the missing information introduces uncertainty in the estimation of the annual amount of social assistance.

Significant issues arise from the annual level of the data because this conflicts with social assistance legislation which determines that social assistance is granted typically on a monthly basis. A report by Tanhua and Kiuru (2020) has described the duration of social assistance receipt in 2019. The study reports that most social assistance recipients receive the benefit for a short period, typically around 1–3 months. In 2019, this group formed 39.4% of all of the basic social assistance recipients, whereas those who received the benefit for a long period, 10–12 months, formed 27.6% of the recipients. On average basic social assistance was granted for 5.8 months, supplementary social assistance for 2.1 months and preventative for 1.9 months in 2019. Around 55% of the total basic social assistance expenditure was granted for recipients receiving the aid for 10–12 months and 12% of the total expenditure was granted for those receiving the aid for 1–3 months. Monthly fluctuations in the need for social assistance are difficult to detect from the annual data because a family who has received social assistance for 1 to 3 months might be earning a higher income annually. The use of annual data could cause bias such that the number of social assistance recipients and the total expenditure on social assistance are underestimated.

Further, the definition of the units of analysis as house-dwelling units does not align with the definition of a family in the Social Assistance Act. This means, for example, that communal households are undetected, although units considered as separate families may live in the same apartment and therefore are required to apply for social assistance separately. This could cause bias, such that the number of families receiving social assistance annually might be underestimated because house-dwelling units tend to form larger groups than a family.

Moreover, the SISU model does not consider uncertainty in take-up behaviour. Non-take-up occurs when a family that would be eligible for basic social assistance does not apply for the benefit (Kuivalainen, 2007). However, the SISU model assumes every potential basic social assistance recipient will take up the benefit. Kuivalainen (2007) investigated non-take-up in Finland in 2005. Some postulated reasons why non-take-up might occur are that the family does not know they would be eligible, a fear of being stigmatized, or the effort in collecting all of the documentation needed for the application. Detection of non-take-up is difficult because the non-take-up rate cannot be reliably estimated from register data sets because often there is missing information about necessary variables, such as wealth. Moreover, questionnaires on non-take-up behaviour have both issues with reliability and missing data. A family's responses could be affected by potentially limited knowledge of their eligibility or social desirability bias could affect what the families disclose, and there is doubt about to what extent the sample is representative of the potential basic social assistance recipients. However, assuming that all potential social assistance recipients take up the benefit could potentially cause bias such that the number of social assistance recipients and the total social assistance expenditure are overestimated.

Finally, the social assistance estimates are affected by possible errors in the other SISU submodels. Other social benefits are calculated before social assistance and these estimates are used in part to calculate the social assistance estimates. Therefore, potential issues with non-take-up behaviour or error in the estimation of the other benefits trickle down to the social assistance estimates. Given that the estimation of error in the other submodels is out of the scope of this study, it is difficult to assess the significance of this error for the social assistance estimates.

3 Social assistance recipients

The following section describes the attributes associated with social assistance recipients in previous literature. This section is organised such that variables relevant from the perspective of the social assistance legislation are presented first. Then, additional variables associated with financial difficulties in previous literature are presented. The findings are compared and contrasted with several corresponding variables in the 2019 register data set. More detailed descriptions of described variables are presented in Appendix Table A2. However, the following review concentrates only on the most relevant predictors due to the wealth of possible predictors available. Note that the continuous variables have been categorised for visual presentation, while otherwise they are treated as continuous.

For the 2019 register data set, a *social assistance recipient* was defined as a house-dwelling unit whose annual social assistance was 1 euro or higher after pos-

sible repayment of the basic social assistance. The *annual social assistance* was defined as the total of annual basic, supplementary and preventative social assistance. The definition was created to investigate the annual amount of social assistance after possible repayments of basic social assistance. However, due to these possible repayments, the annual social assistance could be negative or unrealistically low.

First, according to the social assistance legislation (see Section 1), the eligibility or the monthly social assistance is determined by the applicant's main activity, age, family structure, and income, comprising of for example salary, entrepreneurial income and social benefits. Most often the literature focuses on social assistance recipient status (Ahola and Hiilamo, 2013; Jauhiainen and Korpela, 2019; Tanhua and Kiuru, 2020), or for example how the social assistance is spent (Ahola and Hiilamo, 2013) or what type of income or expenditure the households receiving social assistance have (Ahola and Hiilamo, 2013; Jauhiainen and Korpela, 2019; Tanhua and Kiuru, 2020). The focus is less often on how the amount of social assistance is distributed in groups defined by family characteristics most likely because the amount of basic social assistance is granted based on the income of the family and the family structure as defined by the social assistance legislation (see Section 1). Therefore, the empirical distribution of the annual social assistance according to household attributes is supplemented through the descriptions of house-dwelling units in the 2019 register data set.

It has been previously reported that social assistance receipt varies according to the main activity of the family (Jauhiainen and Korpela, 2019). Typically the reference person of the household is unemployed or laid-off (Tanhua and Kiuru, 2020; Jauhiainen and Korpela, 2019; Ahola and Hiilamo, 2013). For example, out of the social assistance recipients, around 70% were unemployed in 2017–2019 (Jauhiainen and Korpela, 2019; Tanhua and Kiuru, 2020). Both students and persons on a pension together formed fewer than 10% of social assistance recipients in 2017–2019 (Tanhua and Kiuru, 2020; Jauhiainen and Korpela, 2019), and employed persons formed around 6–8% of recipients between 2017 and 2019 (Tanhua and Kiuru, 2020). Notably, disability pensioner status and a long-term illness have been associated with social assistance receipt (Kauppinen et al., 2014; Ahola and Hiilamo, 2013; Vaalavuo, 2016). For example Vaalavuo (2016) investigated families' public health care use and social assistance receipt between 2015–2011. The study suggests that social assistance recipients use public health care more frequently than non-recipients both before and after the first receipt of social assistance. In brief, the main activity of the family's reference person of a family receiving social assistance is typically unemployed and health problems are associated with financial difficulties.

In the 2019 register data set, it appears that the annual social assistance varies

according to the main activity. This data set includes two variables which describe the main activity of a family. Figure 3i visualizes the main activity of the reference person on the last day of the year. The largest proportion of social assistance recipients are long-term unemployed and other non-working (other) groups. Social assistance recipients are less likely to be retired (on pension) or students, which conforms with the previous literature. Similarly, the long-term unemployed and non-working (other) groups receive the highest amounts of social assistance on average. An interesting pattern is that the employed group has the second largest proportion of social assistance recipients after the long-term unemployed, but together with the retired persons, receive the lowest amounts of social assistance on average. In effect, the main activity of the family's reference person is an important predictor of both the social assistance receipt and the amount of annual social assistance.

Further, previous reports suggest the incidence of social assistance varies depending on family structure (Tanhua and Kiuru, 2020; Jauhiainen and Korpela, 2019; Ahola and Hiilamo, 2013). Notably, lone-dwellers form the largest group and it has been reported that between 2008–2010 lone-dwellers formed the largest group in Helsinki (Ahola and Hiilamo, 2013) and 2017–2019, 74% of social assistance recipients lived in a single-person family (Tanhua and Kiuru, 2020; Jauhiainen and Korpela, 2019). Another significant group is single-parent families who formed the second largest group of social assistance recipients in 2017–2019 (Tanhua and Kiuru, 2020; Jauhiainen and Korpela, 2019). In contrast, families consisting of either couples with children or couples without children receive social assistance less frequently (Ahola and Hiilamo, 2013; Tanhua and Kiuru, 2020; Jauhiainen and Korpela, 2019).

According to the social assistance legislation, the family structure in part determines the basic social assistance granted to the applicant. A report by Tanhua and Kiuru (2020) described the basic social assistance according to family structure in 2018 and 2019. The findings suggest that the households with children including single-parent families and couples with children received the greatest amounts of social assistance per month on average. The lowest amounts were granted to single men and the second lowest to single women on average.

The register data set conforms to the finding that lone-dwellers (single) most often receive social assistance compared to other house-dwelling unit structures (Figure 3ii). However, the second largest group in the 2019 register data set is the group denoted as "other" in Figure 3ii, which consists of house-dwelling units where there are more than two adults or more than two adults and children. The single-parent house-dwelling units form the third largest group (Figure 3ii) and the lowest proportion of social assistance recipients is in the two-adult group. Further, the median social assistance amount and the interquartile range are the

highest for single-parent families with children whereas the other groups receive similar amounts of social assistance on average (Figure 3ii). Therefore, the family structure is likely to aid in the prediction of both the probability of observing social assistance receipt as well as the annual social assistance amount.

The amount of basic social assistance granted is determined by the applicant's different sources of income, such as salary, entrepreneurial income, wealth and social benefits. Tanhua and Kiuru (2020) summarised the different sources of income the social assistance recipients typically had in 2017–2019. Social assistance recipients tend to rarely receive salary or entrepreneurial income but receive other social benefits rather often. The proportions of social assistance recipients earning a salary from 2017 to 2019 were 6.5%, 7.4% and 8.4%, respectively. Furthermore, the proportion of social assistance recipients receiving entrepreneurial income was around 0.4%. The most common forms of social benefits were housing allowance of around 79% annually, labour market subsidy of around 44% annually and child allowance of around 19% annually. Additionally, around 8.5% of social assistance recipients had no other income.

The house-dwelling units' salary is connected to social assistance receipt also in the 2019 register data set. First, in Figure 3v it may be observed that the proportion of social assistance recipients appears to decrease as the total annual salary increases. This supports the previous findings, that families eligible for social assistance receive little or no salary. This suggests that earning a small salary might be a good indicator of possible social assistance receipt, and a larger salary might be an indicator of non-receipt. However, due to the annual level, the brief fluctuations in salary possibly causing a need for social assistance may not be detected, which means those receiving social assistance for a brief period are difficult to discriminate based on annual salary alone.

In the case of housing allowance in the 2019 register data set, it appears house-dwelling units receiving larger quantities of housing allowance have a larger proportion of social assistance recipients (Figure 3vi). In terms of the amount, there is a slight positive linear trend between the annual amount of housing allowance and the annual amount of social assistance (Figure 3vi). This suggests that the amount of housing allowance might help to detect social assistance recipients and aid in the prediction of annual social assistance.

Finally, the proportion of social assistance recipients seems to be the highest among house-dwelling units receiving zero days or less than a month of labour market subsidy in the 2019 register data set (Figure 3vi). The families receiving 200–300 days of labour market subsidy form the second largest group. This suggests that either ineligibility for labour market subsidy or long-term unemployment may indicate the need for social assistance.

While age together with the main activity of the family in part determines

the eligibility of the family for basic social assistance, previous studies suggest that young age has been identified as a risk factor for financial difficulties and it often appears to vary together with other risk factors (Ilmakunnas and Moisio, 2019; Ilmakunnas et al., 2015; Tanhua and Kiuru, 2020; Jauhiainen and Korpela, 2019). First, from 2017 to 2018, around half of social assistance recipients were under 35-year-olds, whereas over 65-year-olds formed around 3% of the recipients (Jauhiainen and Korpela, 2019). Further, a study by Ilmakunnas et al. (2015) suggests that employment prospects appear to improve as a function of age. Indication of this was that the incidence of unemployment decreased and the incidence of studying increased as age increased between 18–30 years of age. Moreover, previous studies support that having children in early adulthood is associated both with a risk of receiving social assistance and an increase in the duration of social assistance receipt (Ilmakunnas and Moisio, 2019; Kauppinen et al., 2014).

The 2019 register data set largely supports that young age (Figure 3iv) is associated with social assistance receipt. However, the largest proportion of the social assistance recipients appears to be in the 35–44 age group. However, the 25–35 age group appears to receive the largest annual social assistance amounts on average. Altogether, the age of the reference person appears to predict both the probability of social assistance receipt and annual social assistance, but the effect of age on the probability of the social assistance receipt or annual social assistance is not linear.

In addition, several other factors have been associated with financial difficulties, including the sex of the reference person, housing tenure of the family, education level and immigration status.

The previous research suggests the probability of social assistance varies according to the sex of the applicant. For example, it has been consistently reported that single men tend to receive social assistance more frequently than single women (Ahola and Hiilamo, 2013; Tanhua and Kiuru, 2020). For example, Tanhua and Kiuru (2020) report that single men formed 45% of basic social assistance recipients, and single women formed 30%. In the 2019 register data set, families, where the reference person is male, appear to both receive social assistance slightly more often and receive slightly larger annual amounts of social assistance, but the difference does not appear practically significant (Figure 3iii). These findings suggest that the sex of the reference person may be more significant in interaction with other characteristics, such as family structure identified in previous literature.

Some other relevant factors not visualised in Figure 3 that have been linked to social assistance receipt are housing tenure being rented dwelling, low education level, and immigration status. First, between 2017 and 2018 over 90% of the social assistance recipients lived in a rented apartment (Jauhiainen and Korpela, 2019). Additionally, the number of years spent in education has been associated

with social assistance receipt (Ilmakunnas et al., 2015; Kauppinen et al., 2014). For example, a study by Kauppinen et al. (2014) suggests that between 3–7 years of studying in higher education, a unit increase in student years was associated with a decreased probability of social assistance receipt. Further, immigration status has been associated with social assistance receipt (Ilmakunnas and Moisio, 2019; Ilmakunnas et al., 2015; Ahola and Hiilamo, 2013; Jauhiainen and Korpela, 2019). According to statistics between 2017 and 2019, Finnish citizens formed 84% of social assistance recipients, whereas the second largest group (8%) was from refugee countries (e.g. Afghanistan, Iran) (Jauhiainen and Korpela, 2019).

Descriptive statistics of the housing tenure, education level and immigration status from the 2019 register data are provided in Appendix Table A1. The register data supports that the most common housing tenure is rented dwelling among social assistance recipients, and further suggests that these house-dwelling units receive the highest amounts of social assistance annually on average. Further, the register data also supports that the education level of the reference person is associated with social assistance receipt. The reference person's education level is often a comprehensive school or upper secondary school level among those house-dwelling units that receive social assistance. The mean annual social assistance is the highest among the comprehensive school group. Finally, the register data suggests that most often the reference person of the house-dwelling units which receive social assistance does not have an immigrant status. However, the register data suggest that the house-dwelling units whose reference person has immigrant status receive higher amounts of annual social assistance on average.

4 Models for semicontinuous outcomes

The empirical distribution of annual social assistance is a continuous distribution that is positive and right-skewed and is characterized by a large proportion of zero observations (Figure 1). Outcomes containing both zero and positive continuous values may be called *semicontinuous* (Neelon et al., 2016), because the definition of continuous random variables does not permit adding probability mass to a single point.

Let Y_i denote the annual amount of social assistance received by family i. In the data-generating process, a low-income family has to decide whether to apply for basic social assistance from the National Insurance of Finland. In case the low-income family i would be eligible for basic social assistance, but chooses not to apply for basic social assistance, the value $Y_i = 0$ is observed. In a data-generating process that takes into account non-take-up behaviour, the non-recipient status may originate from one or two potential sources.

First, an occurrence of a non-recipient status may be considered to arise either

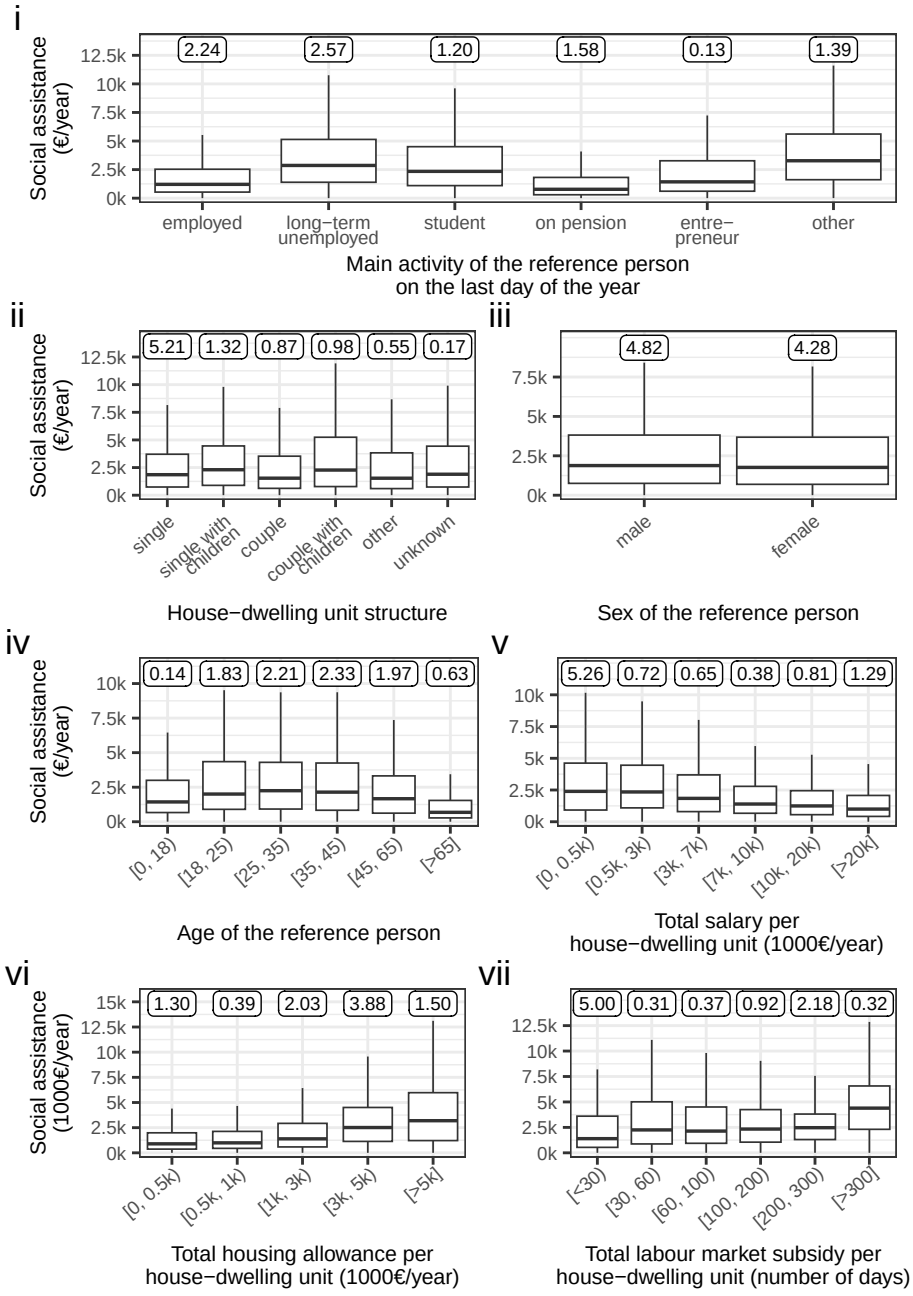

Figure 3: Empirical distributions of the annual social assistance in the 2019 register data set in groups defined by (i) the main activity of the reference person, (ii) the house-dwelling unit structure, (iii) the sex of the reference person, (iv) the age of the reference person, (v) the total annual salary of the house-dwelling unit, (vi) the total annual housing allowance, and (vii) the total annual labour market subsidy. A label on top of each boxplot is a percentage of social assistance recipients within the group in the 2019 register data set. Observations outside 1.5 times the interquartile range are not visualized due to data privacy guidelines.

from a structural source or a non-take-up source (Figure 4). Structural source means those families who do not need social assistance, and non-take-up source means those families who are low-income but do not apply for social assistance. Let Z_i denote the process that determines the family's eligibility status for either basic, supplementary or preventative social assistance for a family i. If they are not eligible non-recipient status $Y_i = 0$ occurs with probability 1. If they are eligible, the second process describes the decision-making regarding take up and the annual amount of basic, preventative and supplementary social assistance Y_i. Let Y_i^* denote the amount of annual social assistance the family i would receive if they choose to apply for the social assistance given that the family is eligible for social assistance. The distribution of social assistance is determined by a detection limit T_i and a parameter vector ω, where T_i represents individual decision-making regarding take-up. If the family is eligible but chooses not to take up the benefit, the potential amount of social assistance is not detected $Y_i^* \leq T_i$ and the value $Y_i = 0$ corresponding to a non-recipient status is observed. If the family chooses to take up the benefit $Y_i^* > T_i$, the potential annual amount of social assistance $Y_i = Y_i^*$ is observed.

Alternatively, an occurrence of a non-recipient status may be considered to arise from one source. In this case, the data-generating process is characterized by two processes (Figure 5) and the decision-making regarding take-up is combined with the eligibility governing process. In other words, the data-generating process simplifies such that the first process Z_i determines the eligibility of the family i for either basic, supplementary or preventative social assistance and the decision to take up the benefit. If $Z_i = 0$, the house-dwelling unit does not receive any annual social assistance, meaning $Y_i = 0$ with probability 1. The second process determines the annual amount of basic, preventative and supplementary social assistance Y_i. If $Z_i = 1$, the house-dwelling unit receives social assistance such that $Y_i > 0$, where their respective distribution for annual social assistance amount is determined by the parameter vector ω.

Whereas the SISU model follows the deterministic data-generating process governed by the Social Assistance Act, a statistical model enables the estimation of the uncertainty arising from the take-up behaviour by estimating the probability that $Y_i = 0$ is observed. In order to capture the data-generating process of both zeros and positive values, a typical option is to model the outcome as a mixture of a degenerate distribution at zero and a base distribution with positive support (Aitchison, 1955). However, as described above, the occurrence of non-recipient status may be considered to have either one or two sources.

From a more general perspective, the above data-generating process may be described as akin to a mixture model, where the mixture components follow different distributions. The eligibility-governing process estimates the mixing probabilities,

and the two mixture components are a discrete point mass at zero for non-recipients or a probability density function for the social assistance recipients. This type of mixture model where one of the mixture components is a point mass is a general strategy to model semicontinuous outcomes, and it has been widely applied to other areas of study such as health services research (Neelon et al., 2015; Cooper et al., 2003), medicine (Moulton and Halsey, 1995) and biology (Hyndman and Grunwald, 2000). Therefore, this mixture modelling framework provides a general approach outside of social assistance and could be applicable to modelling other forms of expenditure, including other social benefits.

Next, two options for modelling semicontinuous outcomes are introduced as possible variations for modelling the distribution of social assistance. Then the choice is presented and an extension of the model with generalised linear modelling is described.

4.1 Zero-inflated Tobit model

The standard Tobit model assumes all observations y_i are from the same base distribution and observations falling below a detection limit $L \in \mathbb{R}^+$ become censored and receive the value zero (Liu et al., 2019; Neelon et al., 2016). The zero-inflated Tobit model gives up the assumption of the common base distribution by adding another zero-generating process (Liu et al., 2019; Neelon et al., 2016).

Let n denote the number of observations, and let $T = (T_1, \ldots, T_n)$ be a vector of detection limits where $T_i \in \mathbb{R}^+$. Let $Z = (Z_1, \ldots, Z_n)$ be a vector of independent indicator variables where $Z_i \sim \text{Bernoulli}(\pi_i)$, and let $Y = (Y_1, \ldots, Y_n)$ be a vector of random variables such that

$$(Y_i \mid Z_i = z_i, T_i = t_i, \omega) \sim \begin{cases} c_0 & \text{when } z_i = 0, \\ c_1(y_i \mid t_i, \omega) & \text{when } z_i = 1. \end{cases}$$

The indicator variable Z_i is latent and determines whether Y_i is from the degenerate distribution at zero or the truncated base distribution. The probability mass function c_0 is such that if $z_i = 0$ then $y_i = 0$ with probability 1. The base distribution $c_1(y_i \mid t_i, \omega)$ is a continuous distribution that is constrained to an interval (T_i, ∞) and it has positive support and a parameter vector ω. In the context of modelling the data-generating process of social assistance, the limit T_i may be considered as a latent random variable, but the model framework would allow the limit to be a known constant. If $Z_i = 1$, then there is a latent Y_i^* which represents the potential value of Y_i for the observation i. If the potential value falls below the detection limit $Y_i^* \leq T_i$, the value Y_i^* becomes censored and $Y_i = 0$ is observed. If the potential value is greater than the detection limit $Y_i^* > T_i$, the value $Y_i = Y_i^*$ is observed. Therefore, a zero from the model can be observed either due to a structural source governed by the process Z_i or due to falling below the detection

limit T_i and becoming censored.

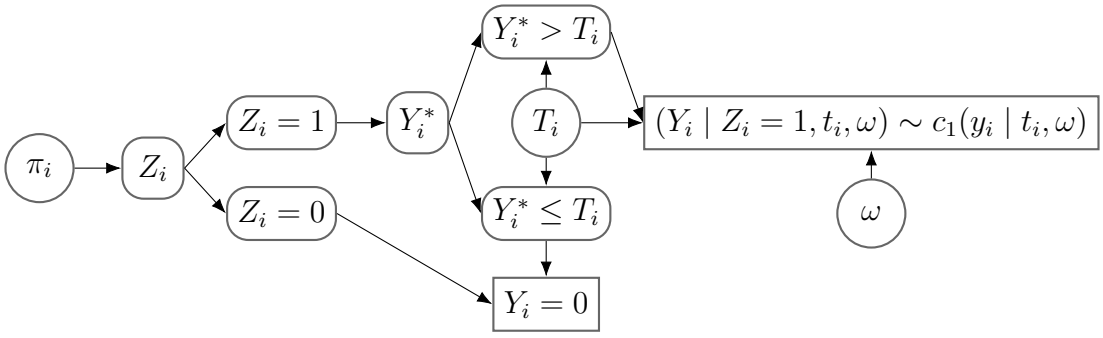

Figure 4: A directed acyclic graph (DAG) which visualizes the data-generating process of the annual social assistance according to the zero-inflated Tobit model. The circles represent parameters, rectangles without rounded corners represent observations and the rectangles with rounded corners represent unobserved variables. Z_i is a latent process that determines the eligibility of the family i for social assistance and $Z_i \sim \text{Bernoulli}(\pi_i)$. Y_i^* represents the potential value of Y_i, and T_i denotes an individual detection limit, which describes the decision-making of the applicant whether to take up the benefit. If the family chooses not to take up social assistance such that $Y_i^* \leq T_i$, then Y_i^* becomes censored and $Y_i = 0$ is observed. If the family chooses to take up the benefit such that $Y_i^* > T_i$, the potential value is observed as $Y_i = Y_i^*$. Y_i follows a continuous, positive distribution truncated at T_i with the parameter vector ω.

4.2 Standard two-part model

The two-part model for semicontinuous outcomes (Liu et al., 2019; Neelon et al., 2016; Aitchison, 1955) is like a zero-inflated Tobit model without censoring. However, the joint probability of eligibility and take-up is determined through a single process Z_i.

Let $Z = (Z_1, ..., Z_n)$ be a vector of independent indicator variables where $Z_i \sim \text{Bernoulli}(\pi_i)$, and let $Y = (Y_1, ..., Y_n)$ be a vector of random variables such that

$$(Y_i \mid Z_i = z_i, \omega) \sim \begin{cases} c_0 & \text{when } z_i = 0, \\ c_1(y_i \mid \omega) & \text{when } z_i = 1 \end{cases}$$

where c_0 refers to a degenerate distribution at zero, where if $z_i = 0$ then $y_i = 0$ with probability 1. The values z_i of the indicator variable are observed. The base distribution c_1 is a continuous probability distribution with positive support and parameter vector ω. Typical choices for the base distribution include log-normal and gamma distributions (Neelon et al., 2016; Liu et al., 2019). Therefore the probability density function for $Y_i \mid Z_i = 1, \omega$ is $c_1(y_i \mid \omega)$. Therefore, the joint

two-part model for Y_i and Z_i is

$$p(y_i, z_i \mid \omega, \pi_i) = p(z_i \mid \pi_i)p(y_i \mid z_i, \omega) = \begin{cases} 1 - \pi_i & \text{when } z_i = 0, \\ \pi_i \cdot c_1(y_i \mid \omega) & \text{when } z_i = 1 \end{cases} \tag{4.1}$$

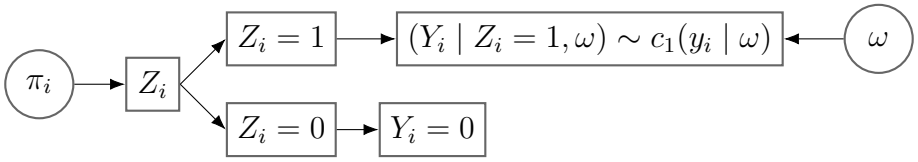

Figure 5: A directed acyclic graph (DAG) for the standard two-part model. The circles represent parameters and the rectangles represent observations. Z_i is a process that determines the status of the social assistance receipt, and $Z_i \sim \text{Bernoulli}(\pi_i)$. Y_i is a process that determines the annual social assistance given Z_i. Y_i follows a continuous, positive and right-skewed distribution with the parameter vector ω.

The zero-inflated Tobit model describes the true data-generating process more realistically in comparison to the standard two-part model because, in the standard two-part model, those not in need of social assistance cannot be differentiated from those choosing not to take up the benefit. However, the occurrence of non-take-up cannot be reliably identified (Kuivalainen, 2007). Therefore, the receipt of social assistance was chosen to be modelled as a standard two-part model.

In the following sections, a model for the take-up of social assistance is referred to as the *binary component*, and the model for the amount of social assistance on the condition of observing take-up is referred to as the *continuous component*.

4.3 Extending the two-part model with regression

Because the SISU model's estimation method adheres to the rules of social assistance legislation, it is restricted to use variables that are relevant to the social assistance legislation to predict the social assistance receipt and the amount of annual social assistance. However, regression analysis enables the use of other risk factors associated with financial difficulties available in the register data set to predict the social assistance receipt and the annual amount of the benefit. In generalised linear modelling, a quantity (e.g. expected value) of a conditional distribution of the response variable is modelled using a linear predictor and a link function. The linear predictor is a linear combination of covariates and their respective regression parameters. The regression parameters are estimated and they aim to describe the relationship of the covariate to the quantity of interest. The standard two-part model may be extended with generalised linear modelling

(Wood, 2017; Fahrmeir et al., 1994) of the binary and continuous components as follows

$$g_0(\pi_i) = g_0(\mathbb{E}[Z_i \mid \pi_i]) = v_i^\top \gamma,$$
$$g_1(\phi_i) = x_i^\top \nu,$$

where $g_0(\cdot)$ and $g_1(\cdot)$ are link functions, π_i is the expected value of a Bernoulli distribution, ϕ_i is a quantity of a continuous distribution, v_i and x_i are covariate vectors, and their respective regression parameter vectors are γ and ν. The symbol $\top$ denotes transpose. For the binary model, the link function $g_0(\cdot)$ could be for example logit or probit link, and for the base distribution the link function $g_1(\cdot)$ could be for example an identity or log link.

In the context of modelling social assistance, the two components of the model have an interpretation. The binary component models the probability π_i to take up social assistance for the house-dwelling unit i, and the continuous component models the distribution of annual social assistance Y_i for the house-dwelling unit i on the condition of social assistance receipt. However, it is important to note that the regression coefficients ν of the continuous part are interpreted on the condition that $Y_i > 0$. Furthermore, the quantity ϕ_i could represent an attribute of the distribution other than the expected value. Therefore, the interpretation of the regression coefficients depends on the attribute of the distribution that ϕ_i describes.

The response variable of the binary component is an indicator variable of whether a house-dwelling unit is a social assistance recipient. The continuous component uses the annual social assistance as the response variable. This further extends upon the SISU model, which only estimates basic social assistance.

5 Model estimation and model selection

In this section, the Bayesian approach to the modelling problem is described first. In the second part, the model selection strategy using an external validation procedure and selection criteria is described. The final section describes which model alternatives were considered as part of the sensitivity analysis.

5.1 Bayesian approach

The Bayesian method was chosen because it was of interest to consider the parameter uncertainty. For example limitations in the data set (see Section 2.4 for details) and non-take-up behaviour introduce uncertainty into the classification and the predictions of the annual social assistance. The use of a Bayesian approach allows the incorporation of parameter uncertainty into the predictions.

This approach enables quantification of the resulting uncertainty as probabilities and facilitates the simulation of distributions for predicted quantities of interest, such as annual social assistance. This provides more conservative estimates in the sense that the variability in the parameter estimates is reflected in predictions, which means the quantities of interest are described by distributions rather than being summarised to point estimates as in the SISU model.

The two components of the model were estimated separately. In equation 4.1 it may be observed that the binary and the continuous components do not have common parameters, which means the prior distributions for the subcomponents may be chosen to be independent. As a result, the posterior distribution has two independent components which can be estimated separately. As a continuation, the model selection was conducted separately for each component.

Let θ denote the parameters of the posterior distribution and y denote the observed data. Samples of θ from the posterior distribution $p(\theta \mid y)$ were simulated using Markov chain Monte Carlo (MCMC). In MCMC, samples of θ are simulated from a Markov chain $\theta^{(s)}$, of which stationary distribution is the posterior distribution $p(\theta \mid y)$ (Gelman et al., 2013). Let $s > 0$ denote the iteration index. Under certain conditions, it may be shown that as the number of iterations increases, the distribution of $\theta^{(s)}$ converges to the stationary distribution, which is independent of the initial values $\theta^{(0)}$ (Gelman et al., 2013). The early iterations of the algorithm are discarded because the chain might not have converged to the stationary distribution. These early iterations are called warm-up iterations. The MCMC algorithm used to simulate the posterior samples was the No-U-Turn Sampler (NUTS, Hoffman et al., 2014) which is an extension of Hamiltonian Monte Carlo (HMC) (Neal, 2011; Betancourt, 2017).

The use of MCMC requires investigation of whether the simulated chains have converged to the stationary distribution. A large number of divergent transitions and potential scale reduction factor for split chains $\hat{R}$ values greater than 1.01 were considered as indicators for non-convergence (Vehtari et al., 2021). The $\hat{R}$ statistic compares the variance within the chains to the variance between the chains for quantities of interest, such as the posterior mean of the model parameters (Vehtari et al., 2021). The $\hat{R}$ statistic is calculated for chains split in half because this helps to identify non-stationary patterns where the chain gradually increases and decreases. The chosen $\hat{R}$ threshold was selected as a heuristic rule to aid in decision-making, but other thresholds are also used in the literature (cf. Gelman et al., 2013). Further, divergent transitions are a diagnostic tool for HMC which may signal that some area of the posterior distribution was not explored sufficiently (Betancourt, 2016; Livingstone et al., 2019). For example, it has been suggested that it is more difficult for HMC to explore areas of the posterior that have strong curvature or heavy tails (Livingstone et al., 2019). Due to a large number of

parameters it was considered too cumbersome to plot and interpret trace plots for each parameter in each model.

The samples drawn from MCMC are serially correlated. Therefore, the effect of the autocorrelation in the chains may be assessed using the effective sample size (ESS) measures. The ESS measures aim to describe the number of independent posterior samples represented by the simulated dependent posterior samples. These measures are calculated after discarding the warm-up iterations. In the analysis, bulk-ESS and tail-ESS measures were considered (Vehtari et al., 2021). In simple terms, the bulk-ESS assesses sampling efficiency at the centre of the posterior distribution and tail-ESS measures the minimum of the ESS at the 5% and 95% quantiles to assess sampling efficiency in the tails of the posterior distribution (Vehtari et al., 2021). Let m denote the number of chains. The effective sample sizes lower than $m \cdot 10$ were considered as an indicator of high autocorrelation within the chains (Gelman et al., 2013). This threshold was selected as a heuristic rule to aid in decision-making, but other threshold choices could also be justified (cf. Vehtari et al., 2021).

Weakly informative prior distributions were used. A weakly informative prior distribution is a proper distribution which has an intentionally larger spread than the available prior information would allow (Gelman et al., 2013). However, given the large sample size, it is to be expected that the choice of the prior distribution does not have a critical effect on the posterior distribution. In general terms, under certain regularity conditions, as the sample size increases the posterior distribution of the parameter vector θ approaches a multivariate normal distribution (Gelman et al., 2013). It follows that as the sample size increases, the relative contribution of the likelihood function dominates the prior distribution in the determination of the posterior. Further details on the asymptotic theory are presented for example in (Gelman et al., 2013). Therefore, other modelling choices, such as the choice of the base distribution may have a greater effect on the modelling results than the choice of the prior distribution. In effect, the influence of the prior distribution alternatives was not considered.

Stan implementation of NUTS was used to estimate posterior distributions (Stan Development Team, 2022). Stan platform was used through the `brms` (Bürkner, 2017) package in `R` (R Core Team, 2023). The `brms` package calls the `rstan` package (Stan Development Team, 2023), which is an `R` interface for using Stan (Stan Development Team, 2022). For each simulation, 3 chains and 2000 sampling iterations were used where the first 1000 iterations were warm-up iterations. The initial values of the parameters were randomly generated (Stan Development Team, 2023). Additionally, all continuous predictors were standardized to improve computation performance. If an indication of non-convergence or high autocorrelation between the chains was observed, the number of chains was increased to 4

or modification of the prior distribution was attempted.

5.2 Model selection through external validation

The wealth of available data prompted opportunities for external validation of the model predictions and concerns for both computation time and memory storage requirements. First, simulation of a posterior distribution may be time-consuming for a large data set using MCMC. Second, the handling and analysis of the register data set were restricted to a remote access system with a limited amount of memory and computational resources available for one user. Therefore, it was considered reasonable to simulate the posterior distribution with a smaller sample size than the full register data set, because it would enable the testing of more model alternatives in a shorter time and thus allow for a broader sensitivity analysis. Second, partitioning the data set into training, validation and testing data sets would provide an opportunity for external validation as a method for model performance evaluation and model selection, while avoiding overfitting the model (Gelman et al., 2013).

The data set was partitioned to enable external validation of the model subcomponents and the full two-part models. First, using simple random sampling, 20% (86592 observations) of the register data set was assigned to the test data set of the full two-part model. The remaining 80%, the non-test set, was used as a training data set for the full two-part model as well as further partitioned into training and validation data sets for the development of the binary and continuous components. The training data set and validation data set were chosen for the binary and continuous components independently. For the continuous component, the training and validation data sets were sampled from the non-test set observations with positive annual social assistance. Of these observations, 80% (25217 observations) were assigned via simple random sampling to the training data set and the remaining 20% (6304 observations) were assigned to the validation data set. For the binary component, the same size training data set (25217 observations) and validation data set (6304 observations) as for the continuous component were chosen using simple random sampling from the non-test set.

To evaluate the performance of the developed models for future observations, predicted values are simulated. In the Bayesian framework, a typical method to generate predictions is to simulate observations from a posterior predictive distribution. Let y denote an observed data set, and let $\tilde{y}$ denote an unobserved future observation from the same data-generating process. Let θ denote the parameter vector of the data-generating distribution of y. The posterior predictive distribution is the distribution of an observation $\tilde{y}$ given the observed data set y (Gelman

et al., 2013) and it is defined as

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta.$$

In this study, new observations were simulated either for the binary component, the continuous component or the full two-part model. During the development of the binary and continuous component candidates for the two-part model, observations were simulated for these submodels separately. For the continuous component, the predicted annual social assistance values were simulated from the posterior predictive distribution given the parameter vector ω. For the binary component, the predicted social assistance receipt statuses were simulated either from the posterior predictive distribution given the parameter π_i or using one of two classification limit methods. See Section 5.4.3 for details on how the receipt statuses were simulated using a classification limit. Let s denote the index of a posterior sample, where $s \in \{1, \ldots, S\}$ and S denote the number of posterior samples. Let the term *posterior predictive* (PP) *classifier* refer to the classification method where the social assistance receipt statuses are simulated directly from the posterior predictive distribution for each posterior sample $\pi_i^{(s)}$.

The full two-part model estimates were generated by first simulating predictions from the binary component either using the PP classifier or one of two classification limit methods. Then, if the observation was predicted to be a social assistance recipient, the annual social assistance was simulated from the posterior predictive distribution of the continuous component. The posterior predictive observations for the test data set observations were simulated by using posterior samples simulated with the training data set and the test data set covariates. For the full two-part model, the following selection criteria were simulated using each of the three classification methods where applicable.

5.3 Selection criteria

The selection criteria were selected to assess the classification accuracy of social assistance recipient status and the predictive accuracy of the annual social assistance received. The selection criteria were calculated for both the test data set and the training data set. For the test data set the criteria were calculated for model selection, whereas for the training data set the criteria were calculated to check the sensibility of the predictions. In this section, the log pointwise predictive density is defined first because it is used in the evaluation of both the binary and continuous component models. In the final two sections, the criteria used specifically to evaluate the continuous component models and the binary component models are presented.

The aim of *the log-posterior predictive density* is to approximate the expected log predictive density (ELPD). The following paragraph uses an adapted version

of the notation used by Gelman et al. (2013). Assuming that a sample θ from the posterior distribution has been simulated using a data set y, and h is the true data-generating distribution, ELPD for a new observation $\tilde{y}_i$ is

$$\text{ELPD} = \int \log p(\tilde{y}_i \mid y) h(\tilde{y}_i) d\tilde{y}_i,$$

which describes the expected value of the log-posterior predictive density for $\tilde{y}_i$ with respect to the true data-generating distribution of $\tilde{y}_i$. The higher the probability density is, the closer the log-posterior predictive density would be to the true data distribution h.

Given the true data distribution h is not known, it is approximated by the test data set y_{test}. To approximate ELPD for the test data set using a model fitted with the training data set, first a posterior distribution is simulated with the training data to get posterior samples $\theta^{(s)}$. Let $y_{\text{test},i}$ denote one observation from the test data set. Then the computed *log pointwise predictive density* (LPPD) for the test data set y_{test} is calculated as

$$\text{LPPD} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_{\text{test},i} \mid \theta^{(s)}) \right)$$

As for the theoretical ELPD, a higher LPPD indicates a better fit for the test data set.

Let M_c denote a model and let M_b denote a model with the highest LPPD among the compared models, where the models within the comparison have been fit with the same data set. Let $\text{LPPD}_{M_{*,i}}$ denote the log posterior predictive density of the observation y_i for model M_*. Therefore, the difference in the log posterior predictive densities for the observation y_i between models M_c and M_b is $\text{LPPD}_{\text{diff},i} = \text{LPPD}_{M_{b,i}} - \text{LPPD}_{M_{c,i}}$. The candidate models were compared by calculating the difference $\text{LPPD}_{\text{diff}} = \sum_{i=1}^{n} \text{LPPD}_{M_{b,i}} - \sum_{i=1}^{n} \text{LPPD}_{M_{c,i}}$ and calculating the standard error of this difference. The estimate of the standard error of $\text{LPPD}_{\text{diff}}$ is defined as in (Sivula et al., 2020) such that

$$\widehat{\text{SE}}_{\text{diff}} = \left(\frac{n}{n-1} \sum_{i=1}^{n} \left(\text{LPPD}_{\text{diff},i} - \frac{1}{n} \sum_{j=1}^{n} \text{LPPD}_{\text{diff},i} \right)^2 \right)^{1/2}$$

If the $\text{LPPD}_{\text{diff}}$ was greater than two standard errors $\widehat{\text{SE}}_{\text{diff}}$, the LPPD value of the model M_c was considered higher. This type of heuristic was selected to aid in decision-making whether models' LPPD values are different. It is based on the assumption that according to the central limit theorem, the distribution of $\text{LPPD}_{\text{diff}}$ approaches normal distribution (Sivula et al., 2020). However, this decision-making may be flawed for example if the normality assumption of $\text{LPPD}_{\text{diff}}$ does not hold or if the model is misspecified (Sivula et al., 2020).

5.3.1 Selection criteria for the continuous component

The predictive accuracy of the continuous component was assessed by the LPPD, the difference to the model with the highest LPPD, the standard error of the total LPPD difference, root mean squared error, the Kolmogorov-Smirnov test statistic, and Bayes-p-values. In the following sections, let y_{rep} denote a sample from the posterior predictive distribution and $y^{(s)}_{\text{rep},i}$ denote a posterior predictive observation for the observation y_i given posterior sample $\omega^{(s)}$.

The *root mean squared error* (RMSE) is a statistic which describes the distance between the observation and its estimate. RMSE is always positive and a zero RMSE value would indicate the observations have been predicted perfectly. In practice, a lower RMSE value indicates a closer fit of the model predictions to the observed values. In the Bayesian framework, the estimate is an observation from the posterior predictive distribution given $\omega^{(s)}$. The RMSE is calculated for a posterior sample $\omega^{(s)}$ as

$$\text{RMSE}^{(s)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y^{(s)}_{\text{rep},i}\right)^2}$$

When calculated for each posterior sample s, the result is a vector of length S comprising of $\text{RMSE}^{(s)}$ values. In this study, the two models were considered different with respect to the RMSE if their 95% credible intervals of the RMSE did not overlap.

Further, the *two-sample Kolmogorov-Smirnov* (KS) *test statistic* d (Schröer and Trenkler, 1995) measures the maximum distance between the empirical distribution functions of two data sets. A two-sided test was chosen, because the aim was to test whether observations from the test data set and simulated observations from the posterior predictive data set given posterior sample s could come from identical distributions. The null hypothesis of the two-sample KS test is that the two samples arise from the same distribution, and a large test statistic value d would indicate that the samples might arise from different distributions (Schröer and Trenkler, 1995). The test statistic for a posterior sample $\omega^{(s)}$ is calculated using formula

$$d^{(s)} = \max_{t\in\mathbb{R}} \mid \hat{F}_y(t) - \hat{F}_{y^{(s)}_{rep}}(t) \mid$$

where $\hat{F}_j(\cdot)$ is the empirical distribution function estimated from the data set j. When calculated for each s, the result is a vector of length S comprising of $d^{(s)}$ test statistic values. Two models were considered different with respect to the KS test if their 95% credible intervals of the KS test statistic d did not overlap. The test was employed using the `ks.test` function of the stats package in `R`, which uses the algorithm developed by Schröer (1991) and Schröer and Trenkler (1995) to estimate the empirical distribution functions.

Moreover, Bayes-p-values for test quantities of interest were calculated to investigate the predictive accuracy of the empirical distribution of the observed data set. Bayes-p-value represents the probability that a test quantity calculated from the posterior predictive sample is more extreme than a test quantity calculated from the observed data set (Gelman et al., 2013). Therefore, Bayes-p-values close to 0.5 would indicate that the model captures the observed test quantity in expectation. Let $\mathbb{I}$ denote an indicator function, and $T(y)$ a test quantity calculated from an observed data set. The Bayes-p-value may be presented as

$$p_{\text{Bayes}} = P\Big(T(y_{\text{rep}}) \geq T(y) \mid y\Big) = \int \mathbb{I}(T(y_{\text{rep}}) \geq T(y)) p(y_{\text{rep}} \mid y) dy_{\text{rep}}$$

This is approximated by first simulating samples $y_{\text{rep}}^{(s)}$ from the posterior predictive distribution for each posterior sample s. Then the Bayes-p-value is calculated using the following formula

$$\hat{p}_{\text{Bayes}} = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\Big(T(y_{\text{rep}}^{(s)}) > T(y)\Big)$$

The chosen test quantities were the mean, total, variance and 99% quantile. The maximum could not be selected due to data privacy guidelines.

In the case when a classification limit was used, Bayes-p-values cannot be calculated according to their definition, because the predicted social assistance statuses are not observations from the posterior predictive distribution (see Section 5.4.3). However, a similar quantity to a Bayes p-value can be calculated, but instead of simulating recipient statuses from the posterior predictive distribution, the predicted statuses are generated as described in Section 5.4.3. To avoid confusion, let the term *predictive p-value* refer to the measure similar to the Bayes-p-value, but where the annual social assistance amounts were simulated using recipient statuses which were generated using a fixed classification limit.

The continuous component was first selected based on LPPD. If no differences between the compared models were found, the models were compared based on RMSE. If no difference was found, the models were compared based on the KS test d statistic. If this was not sufficient to choose the best-performing model, the models with the fewest parameters were identified among those models with the highest LPPD. From these models, the model with the highest LPPD was selected.

5.3.2 Selection criteria for the binary component

The binary component was first selected based on LPPD. In this case, the log-likelihoods were calculated for the social assistance receipt statuses from the test data, utilising the posterior simulations s corresponding to $\pi_i^{(s)}$. If this was not sufficient to choose the best-performing model, the models with the fewest parameters were identified among those models with the highest LPPD. From these

models, the model with the highest LPPD was selected.

The best-performing logistic regression model developed as the binary component was also evaluated by calculating a receiver operating characteristic (ROC) curve (e.g. Metz, 1978). The ROC curve may be plotted for binary classifiers which generate probabilities to belong to a positive class, and it visualises the trade-off between the true positive rate and the false positive rate when the classification limit is varied. In this study, the social assistance receipt statuses were simulated using the classification method described in Section 5.4.3. For each posterior sample s, the true positive rate and the false positive rate were calculated for classification limits from 0 to 1 in steps of 0.01. Then, the mean and 99% credible interval of the false positive rate and the true positive rate were calculated for each classification limit.

Additional statistics used in the evaluation were the misclassification rate, the false negative rate and the false positive rate. The predictions of the social assistance status were simulated using a classification limit and these statistics were calculated for each posterior sample s, and the mean and 95% credible interval of each measure was investigated. However, neither the misclassification rate, false negative rate nor false positive rate were used in the selection of the logistic regression model for the binary component. For a description of how the predictions were simulated using the classification limit and how the impact of a classification limit was investigated, see Section 5.4.3.

5.4 Sensitivity analysis

5.4.1 Predictors

Given the wealth of available predictors, it would be time-consuming to exhaust all possible model alternatives. Therefore, the focus of the predictor selection was to choose a group of predictors considered relevant for the predictive accuracy of house-dwelling-level estimates. Then, a few selected combinations of predictors were altered based on hypotheses of interest.

First, the predictors which were considered relevant were those connected to social assistance through legislation, previous literature, or predictors which might help to correct a known error in the SISU model estimates (Section 2.4). The following predictors were included in each model on these grounds.

Predictors describing the house-dwelling unit qualitatively were included as categorical predictors. The chosen categorical variables were the sex of the reference person, the structure of the house-dwelling unit, the main activity of the reference person, housing tenure, the education level of the reference person, and the immigrant status of the reference person and income decile. Age of the reference person was included as a categorical predictor with levels $[0, 18), [18, 25), [25, 35), [35, 45)$,

$[45, 65)$ and $[65, \infty)$. Additionally, an indicator for a communal household was included as a predictor to potentially capture variation unaccounted for by defining the units of analysis as house-dwelling units. This indicator describes whether persons who are 18 or older and are not the spouse of the reference person live in the same house-dwelling unit.

The included continuous predictors were the salary, capital income, the number of months the reference person was unemployed, and several social benefits. The included social benefits were the number of days the adults (18 years or older) of the house-dwelling unit received labour market subsidy, the number of days the adults of the house-dwelling unit received earnings-related unemployment allowance, the number of days the adults of the house-dwelling unit received basic unemployment allowance, the number of months the reference person received student aid, total annual housing benefits of the house-dwelling unit, and the total annual pension of the house-dwelling unit. Additionally, the sum of the other social benefits was included.

Further, it was of interest to test whether variations in the predictor structure would improve predictive accuracy:

1. Inclusion of either social benefits estimated by the SISU model or observed social benefits in the register data set (2 variants).

2. Inclusion of social assistance estimated by the SISU model or its transformation. One alternative was to categorize the SISU model estimated basic social assistance, and another alternative was to omit the variable (2 variants).

3. Two alternative variables describing the main activity of the reference person. The first measured the main activity on the last day of the year and the second aimed to describe the main activity over the course of a year. Due to a large number of factor levels, the second variant was recategorised to entrepreneur, employed, student, retired and other classes, and the first variant was recategorized otherwise the same but it was possible to create a separate class for house-dwelling units whose reference person had a long-term unemployed status (1 variant).

4. Inclusion of salary on the original scale or a transformation of the annual salary. Two transformations were considered: log transformation and scaling the annual salary with the number of adults (18 years old or older) in the house-dwelling unit (2 variants).

5. Inclusion of interaction terms which account for the house-dwelling unit structure. The tested interaction terms were house-dwelling unit structure and the sex of the reference person, the house-dwelling unit structure and

salary, the house-dwelling unit structure and housing benefits. These interaction terms were included one by one, and then pairwise such that two of the three interaction terms were included simultaneously in one model (6 variants).

After each of these variations, the best-performing candidate was selected using the external validation procedure described in Section 5.2. This candidate was used as the starting point model for the next group of variations. For example, if it was found that the model where the salary was scaled with the number of adults living in the house-dwelling unit had the best performance, this variant of salary was used in interactions in the next predictor variant group. Altogether these variations produced 13 candidate models. The descriptions of selected variables which were used after model selection are provided in Appendix Table A2.

5.4.2 Distribution assumption

First, the link function of the regression used to model the binary component was altered to assess whether the classification results could be improved. Therefore, both logit and complementary log–log link functions were considered. However, the complementary log–log link was left out of consideration due to a high number of divergent transitions. Therefore, the 13 candidate models outlined above were fitted using a logistic regression model.

Second, the distribution assumption of the continuous component was altered to assess whether the predictive accuracy of the annual social assistance could be improved. Several distribution assumptions and parameterisations were considered for the continuous component initially. The initial inspection included fitting the model first with a smaller set of predictors with the training data set, and potentially varying the prior distribution, link function, and parameterisation in order to improve convergence. The considered distributions were exponential, Weibull, gamma, log-normal, generalised gamma, generalised extreme value and generalised Pareto distributions. The generalised gamma and generalised Pareto distributions were not supported by default in `brms` and therefore were implemented by the author. The log-likelihood and random number generation functions were written in `Stan`.

First, the generalised extreme value distribution was left out of consideration, because large $\hat{R}$ values suggested poor convergence and effective sample size measures suggested high autocorrelation within the chains. The most promising initial results were given by Weibull, gamma and generalised gamma distributions with regard to RMSE, KS test and visual inspection of the degree to which the posterior predictive samples resembled the empirical distribution of annual social assistance. Therefore, these three distributions were selected for further consideration.

The Weibull and gamma distributions were parameterised with respect to the expected value and a log link was used. The generalised gamma distribution was parameterised as in Stacy (1962) with adapted notation. Let $\alpha > 0$, $\beta > 0$ and $\delta > 0$. For $y_i > 0$, the used probability density function was

$$f(y_i \mid \alpha, \beta, \delta) = \frac{\delta}{\beta^{\delta\alpha}\Gamma(\alpha)} y_i^{\delta\alpha-1} \exp\Big(- \Big(\frac{y_i}{\beta}\Big)^{\delta}\Big)$$

where $\Gamma(\cdot)$ denotes the gamma function. The parameter β is a scale parameter, meaning it controls the spread of the distribution, and the parameters α and δ are shape parameters (Stacy and Mihram, 1965). The generalised gamma distribution includes Weibull and gamma distributions as special cases (Stacy, 1962). The Weibull distribution follows from the parameterisation $\alpha = 1$ and the shape–scale parameterisation of the gamma distribution follows directly from the parameterisation $\delta = 1$.

Moreover, the impact of including a regression component for both the expected value and the parameter α of the generalised gamma distribution was investigated by fitting the 13 predictor structure alternatives with both parameterisations. In both cases a log link was used. The models were evaluated using the external validation procedure described in Section 5.2. However, the investigation was restricted to generalised gamma distribution where the regression was set to parameter α because the models with this parameterisation produced the most promising results with respect to RMSE, KS test and the degree to which the visualisations of the shape of the posterior predictive samples of the annual social assistance resembled the empirical distribution of the annual social assistance.

Furthermore, regression on another parameter of the base distribution was investigated to test whether extending the regression to both parameters of Weibull and gamma distributions would improve the fit of the model to the empirical distribution of social assistance. However, the improvements in predictive accuracy were modest and therefore this option was not investigated further.

5.4.3 Classification limit

Initial results suggested that the PP classifier produced classification results that were rather modest compared to the SISU model (see Section 6 for comparisons). Therefore, it was of interest to employ another classification method.

Let $\hat{\pi}_i^{(s)}$ denote the probability of social assistance recipient status $\hat{z}_i^{(s)} = 1$ for house-dwelling unit i for posterior sample s. A prediction of social assistance recipient status was generated such that

$$\hat{\hat{z}}_i^{(s)} = \begin{cases} 1 & \text{when } \hat{\pi}_i^{(s)} \geq \ell, \\ 0 & \text{when } \hat{\pi}_i^{(s)} < \ell \end{cases}$$

where $\ell \in [0, 1]$ was a selected classification limit.

There was no specific classification target, other than to obtain a better misclassification rate than the SISU model. This means the choice of the limit ℓ is ambiguous and various classification limits could be satisfactory. Moreover, because of the unbalanced proportions of benefit recipients and non-recipients, a classification limit ℓ other than 0.5 could provide a desirable classification result. Therefore, while the final choice of the limit is left to the user, two exemplary limits ℓ are selected heuristically by varying the classification limit and choosing the most appropriate classification limit based on an optimisation measure.

The considered optimisation measures were the misclassification rate, F-score and Cohen's kappa. However, the F-score was left out of consideration because the choice of the parameter $\beta \in \mathbb{R}$, which tunes the preference over precision or recall, is ambiguous, and the search for an appropriate value of β would have required further computation.

Let $n_c^{(s)}$ denote the number of correctly predicted social assistance receipt statuses for posterior sample s. Let $n_0^{(o)}$, $n_1^{(o)}$, $n_0^{(s)}$, and $n_1^{(s)}$, denote the number of observed non-recipients, the number of observed social assistance recipients, the estimated number of non-recipients for posterior sample s, and the estimated number of social assistance recipients for posterior sample s. Cohen's kappa (Cohen, 1960; Warrens, 2008) is defined as

$$\kappa^{(s)} = \frac{p_o^{(s)} - p_e^{(s)}}{1 - p_e^{(s)}},$$

where

$$p_o^{(s)} = \frac{n_c^{(s)}}{n} \quad \text{and} \quad p_e^{(s)} = \frac{n_0^{(o)}}{n} \cdot \frac{n_0^{(s)}}{n} + \frac{n_1^{(o)}}{n} \cdot \frac{n_1^{(s)}}{n}.$$

The range of Cohen's kappa is $[-1, 1]$ (Warrens, 2008) where $\kappa^{(s)} = 1$ indicates perfect agreement. This means values of $\kappa^{(s)}$ which are positive and close to one are preferred.

The classification limit was selected as follows. Let k denote index of a limit ℓ_k, where $k \in \{1, \ldots, K\}$ and K is the number of limits tested. For each limit ℓ_k the predictions of the social assistance recipient status were simulated as described in the beginning of this section for each posterior sample s. This produced $S \times K$ estimates of a given optimisation measure. Then, a mean of the optimisation measure was calculated over the posterior samples for each limit ℓ_k and the limit which gave the best result regarding the optimisation measure was selected. For the misclassification rate, the lowest mean was used to indicate the best limit and for Cohen's kappa the highest mean was used to indicate the best limit. Let the terms *kappa-optimised* (KO) *classifier* and *misclassification rate-optimised* (MRO) *classifier* refer to the classification methods where the classification is conducted

using a classification limit, and the classification limit was selected by optimising Cohen's kappa or the misclassification rate, respectively.

In selecting the logistic regression model for the binary component, the impact of the classification limit was investigated by testing $K = 6$ possible limits from 0.2 to 0.7 in steps of 0.1. However, the same limit was selected for each model regardless of the predictor structure, given the optimisation measure was fixed. In effect, the variation in the classification limit did not influence the misclassification rate, the false positive rate or the false negative rate. Therefore, these measures were not used to select the best-performing logistic regression model for the binary component.

Therefore, two options for a classification limit were selected based on optimising Cohen's kappa or the misclassification rate for the best-performing logistic regression model that was chosen for the full two-part model. Altogether $K = 100$ limits were tested from 0 to 1 in steps of 0.01. The limit was selected using the training data set of the two-part model and 3000 posterior samples.

In summary, the 13 candidate logistic regression models of the binary component were fitted with a training sample. Then, the 13 candidate models assuming either a Weibull, gamma or generalised gamma distribution were fitted with a training sample. One best-performing logistic regression model and one best-performing model from each of the continuous distribution assumptions were selected using a validation data set for external validation. By combining the best-performing logistic regression model with each of the models for the continuous component, three standard two-part models were formed. Additionally, for each two-part model, the classification was conducted using the PP, KO, and MRO classifiers. The performance of these models with respect to the three alternative classification methods was evaluated using a test data set for external validation. The results of this evaluation are presented next.

6 Results

In this section, the results for the three developed two-part models with respect to three alternative classification methods are compared to each other and the SISU model using selection criteria defined in Section 5.2 where applicable. First, the classification accuracy results with respect to the classification methods are presented, and then the results regarding the amount of annual social assistance are presented. The regression coefficients for each submodel are not further discussed, but the descriptions of the selected covariates are available in Appendix Table A2 and the posterior means and standard errors are available in Appendix Tables A2–A7.

For all of the fitted models, $\hat{R}$ values 1.01 or lower were observed and both

bulk-ESS and tail-ESS values were greater than the threshold value $m \cdot 10 = 30$ where the number of chains $m = 3$. Therefore the Markov chains were considered to have converged for each fitted model. The estimated $\hat{R}$ values and effective sample size values are reported in Appendix Tables A3–A7.

The classification accuracy results depend on the choice of the classification method. The limit selected using the MRO classifier was 0.51 and the limit selected using the KO classifier was 0.33. In the context of conducting the classification using a classification limit, the trade-off between the true positive rate and the false positive rate for the best-performing logistic regression model is plotted in an ROC curve in Figure 6. In general, it may be observed that the ROC curve is far from a diagonal line, which represents a curve that would result from a random guess. There was very little variation in the true positive and false positive rates at each limit, as the 99% uncertainty interval was too narrow to be well discernible in the figure.

Further, the four points in Figure 6 visualise the mean false positive rates and mean true positive rates of the three developed classifiers and the SISU model with respect to the ROC curve. The means were plotted because the variation in these measures was small. The KO and MRO classifiers represent two possible options on the ROC curve. The KO classifier generates a higher true positive rate at the expense of the false positive rate, whereas the MRO classifier generates a lower false positive rate at the expense of the true positive rate. Notably, at the false positive rate of the PP classifier, the SISU model produces a higher true positive rate. Similarly, at the true positive rate of the PP classifier, the SISU model produces a lower false positive rate. However, at the false positive rate of the SISU model, the developed logistic regression model produces a higher true positive rate than the SISU model. Similarly, at the true positive rate of the SISU model, the developed logistic regression model produces a lower false positive rate than the SISU model. Note that for the dummy model, the true positive rate and false positive rate are zero.

The classification accuracy results are presented in Table 1. For the PP classifier, the false positive rate is similar to that of the SISU model, while its false negative rate is higher than that of the SISU model. Furthermore, the PP classifier's misclassification rate is higher than that of the SISU model and similar to that of the dummy model. However, when the classification is conducted using either the KO or MRO classifiers, the misclassification rate is lower than that of the SISU model and that of the dummy model. Moreover, the KO classifier produces a lower false negative rate as well as a lower false positive rate compared to the SISU model. In contrast, the MRO classifier's false negative rate is higher than the false negative rate of the SISU model. Among the candidate two-part models, the model fit with the generalised gamma distribution had the highest LPPD (Table

2). The model fitted with the Weibull distribution had the second highest LPPD and the model fitted with the gamma distribution had the lowest LPPD. Further, the $\text{LPPD}_{\text{diff}}$ values of Weibull and gamma models were two standard errors apart from the generalised gamma

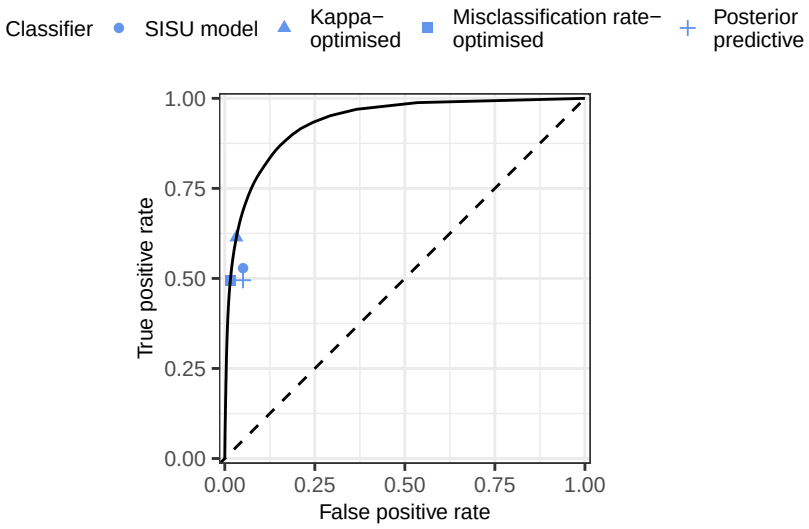

Figure 6: The receiver operating characteristic (ROC) curve of the standard two-part model. The individual points represent the mean false positive rates and the mean true positive rates of the three alternative classification methods, and the corresponding point estimates of the SISU model. The estimates have been calculated using the training data set of the standard two-part model. The 99% credible interval is not plotted because it was too narrow to be well discernible visually.

distribution (Table 2). This suggests the LPPD is higher for the model fitted with the generalised gamma distribution.

The choice of the classification method had a further impact on the predictions generated from the continuous component. The RMSE and KS test d-statistic values of the SISU model and the developed models with respect to the classification methods are presented in Table 3.

In general, the mean RMSE values of the candidate models are smaller than the SISU model estimate (Table 3). The SISU model estimate is not included in the 95% credible intervals when assuming the Weibull or generalised gamma distributions for the continuous component, regardless of the classification method. However, the models using the MRO classifier produced the lowest RMSE estimates on average. Additionally, the SISU model estimate is not included in the 95% credible interval when assuming the gamma distribution and using either KO or MRO classifiers. However, when assuming the gamma distribution and using

Table 1: Mean and 95% credible intervals of the false negative rate, false positive rate and misclassification rate for the standard two-part model according to the classification method, and the corresponding point estimates of the dummy model and the SISU model. The estimates have been calculated using the test data set, and the two-part model the predictions have been simulated using 3000 posterior samples.

Classifier	Statistic	Estimate (%)	95% CI
Dummy			
	Misclassification rate	9.13	
SISU model			
	False negative rate	46.68	
	False positive rate	5.19	
	Misclassification rate	8.78	
Two-part model			
Posterior predictive	False negative rate	50.18	[49.30, 51.11]
	False positive rate	5.05	[4.91, 5.19]
	Misclassification rate	9.17	[9.02, 9.32]
Kappa-optimised	False negative rate	38.29	[37.92, 38.65]
	False positive rate	3.17	[3.09, 3.24]
	Misclassification rate	6.38	[6.33, 6.42]
Misclassification rate-optimised	False negative rate	50.22	[49.86, 50.58]
	False positive rate	1.49	[1.46, 1.53]
	Misclassification rate	5.94	[5.92, 5.97]

the PP classifier, the SISU model estimate is included in the 95% credible interval. Notably, the RMSE values are the lowest for the models fitted with the generalised gamma distribution.

Among the developed models, the distribution assumption had a negligible effect on the KS test d-statistic values (Table 3). The mean d-statistic values are equivalent when the receipt statuses are simulated using either the PP or KO classifiers. Both of these methods produce d-statistic estimates that are lower than the SISU model d-statistic estimate, and the SISU model d-statistic estimate is not included in any of the 95% credible intervals. However, when the predictions are simulated using the MRO classifier, the produced d-statistic estimates are higher on average than the d-statistic estimates of the other classifiers. Moreover, the MRO classifier produces higher d-statistic estimates than that of the SISU model on average, and the SISU model d-statistic is not included in the 95% credible intervals.

Moreover, both the distribution assumption and the choice of classification method impacted the extent to which values simulated from the candidate models

Table 2: LPPD, LPPD differences ($\text{LPPD}_{\text{diff}}$), and standard error of the LPPD differences ($\widehat{\text{SE}}_{\text{diff}}$) of the two-part model with varying continuous distribution assumption (Model) of total annual social assistance. The differences in LPPD have been calculated with respect to the model with the highest LPPD. The estimates have been calculated from posterior predictive samples using the test data set.

Model	LPPD	$\text{LPPD}_{\text{diff}}$	$\widehat{\text{SE}}_{\text{diff}}$
Generalized gamma	-82191.71	0.00	0.00
Weibull	-82473.30	-281.59	50.61
Gamma	-82537.22	-345.51	48.17

resembled the empirical distribution of annual social assistance. The (Figure 7) visualises the distributions of mean, total, variance and 99% quantile in groups defined by the distribution assumption and the KO and MRO classifiers, and the corresponding estimates of the SISU model. In general, the developed model predictions appear to fit the empirical annual social assistance distribution more accurately than the SISU model predictions fit the empirical basic social assistance distribution.

The developed models' mean and total predictions appear insensitive to the distribution assumption but are rather sensitive to the choice of the classification method (Figure 7 i, ii). If the classification was conducted using the KO classifier, the predicted mean and total were slightly overestimated regardless of the distribution assumption. For the KO classifier, the predictive p-values of the mean and total were 1.0 for each of the continuous distributions. However, if the classification was conducted using the MRO classifier, the predicted mean and total were underestimated regardless of the distribution assumption. For the MRO classifier, the predictive p-values of the mean and total were 0.0 for each of the continuous distributions. On the other hand, for the PP classifier, Bayes-p-values of the mean and total were 0.39, 0.04 and 0.04 for the gamma, Weibull and generalised gamma distributions respectively. Furthermore, the mean and total annual basic social assistance point estimates predicted by the SISU model deviated more from their observed values than the mean and total annual social assistance estimates from the developed two-part models deviated from their observed value.

The variance predictions appear to be sensitive to both the distribution assumption and the choice of the classification method (Figure 7 iii). In general, the two-part model fitted with the generalised gamma distribution generated more accurate variance estimates than the two-part models fitted with the Weibull or gamma distributions. For the Weibull and gamma distributions, the predictive p-values for the variance were 1.0 regardless of the classifier. For the generalised gamma distribution, the predictive p-values were 1.0 and 0.02 for the KO and MRO

Table 3: Mean and 95% credible intervals of the root mean squared error (RMSE) and Kolmogorov-Smirnov (KS) test d-statistic for each of the standard two-part models according to the continuous distribution assumption and the classification method, and the corresponding point estimates of the SISU model. The estimates have been calculated using the test data set, and the two-part model the predictions have been simulated using 3000 posterior samples.

Classifier	Criteria	Estimate	95% CI
SISU model			
	RMSE	1426.76	
	KS test d	0.019	
Gamma			
Posterior predictive	RMSE	1330.10	[1267.36, 1432.87]
	KS test d	0.007	[0.006, 0.008]
Kappa-optimised	RMSE	1271.38	[1207.43, 1360.41]
	KS test d	0.006	[0.005, 0.007]
Misclassification rate-optimised	RMSE	1201.63	[1148.36, 1270.96]
	KS test d	0.032	[0.032, 0.033]
Weibull			
Posterior predictive	RMSE	1274.59	[1219.12, 1365.19]
	KS test d	0.006	[0.005, 0.007]
Kappa-optimised	RMSE	1200.08	[1149.06, 1272.05]
	KS test d	0.006	[0.005, 0.007]
Misclassification rate-optimised	RMSE	1133.35	[1091.11, 1185.82]
	KS test d	0.032	[0.032, 0.033]
Generalized gamma			
Posterior predictive	RMSE	1108.43	[1088.47, 1128.74]
	KS test d	0.002	[0.001, 0.003]
Kappa-optimised	RMSE	1026.00	[1006.50, 1045.42]
	KS test d	0.006	[0.005, 0.007]
Misclassification rate-optimised	RMSE	979.07	[961.11, 996.53]
	KS test d	0.032	[0.032, 0.033]

classifiers, respectively. Moreover, for the PP classifier, the Bayes-p-values were 1.0 for the Weibull and gamma distributions and 0.27 for the generalised gamma distribution. Furthermore, the two-part model's variance estimates of the annual social assistance are more accurate on average than the SISU model's variance estimate of the annual basic social assistance.

Finally, the 99% quantile estimates appear to be insensitive to the distribution assumption, but sensitive to the choice of the classification method (Figure 7 iv). Similarly, as for the mean and total estimates, if the predictions were generated using the KO classifier, the two-part models slightly overestimated the empirical 99% quantile. For the KO classifier, the predictive p-values for the gamma, Weibull and generalised gamma distributions were 0.95, 0.36 and 0.97, respectively. However, if the classification was conducted using the MRO classifier, the models slightly underestimated the observed 99% quantile. For the MRO classifier, the predictive p-values were 0.0 for the gamma and Weibull distributions, and 0.01 for the generalised gamma distribution. Furthermore, for the PP classifier, Bayes-p-values were 0.01, 0.0 and 0.0 for the gamma, Weibull and generalised gamma distributions, respectively. Furthermore, the SISU model estimate of the 99% quantile of basic social assistance 99% quantile deviates more from its observed value than the two-part model estimates of the 99% quantile of annual social assistance deviates from their observed value.

7 Discussion

The aim of the modelling was to apply a two-part model to social assistance estimation using the Bayesian framework and to assess whether the developed model would improve the predictive accuracy of the SISU model's annual social assistance estimates. The two-part model utilises a strategy akin to mixture modelling to combine a submodel estimating the probability of observing a social assistance receipt status and a submodel estimating the amount of annual social assistance. Given that the house-dwelling unit is a social assistance recipient, the estimated annual social assistance is positive and otherwise zero. Altogether three two-part model candidates were developed by altering the continuous distribution assumption. The investigated continuous distributions of the annual social assistance were gamma, Weibull or generalised gamma distributions. In addition, three alternative classification methods to predict social assistance receipt were investigated. The initial method was to simulate the receipt statuses using a posterior predictive classifier, which simulated the receipt statuses directly from the posterior predictive distribution. Alternatively, the classification was conducted by using the probabilities of social assistance receipt and choosing the predicted recipient status based on a classification limit. Kappa-optimised classifier was developed by choosing the

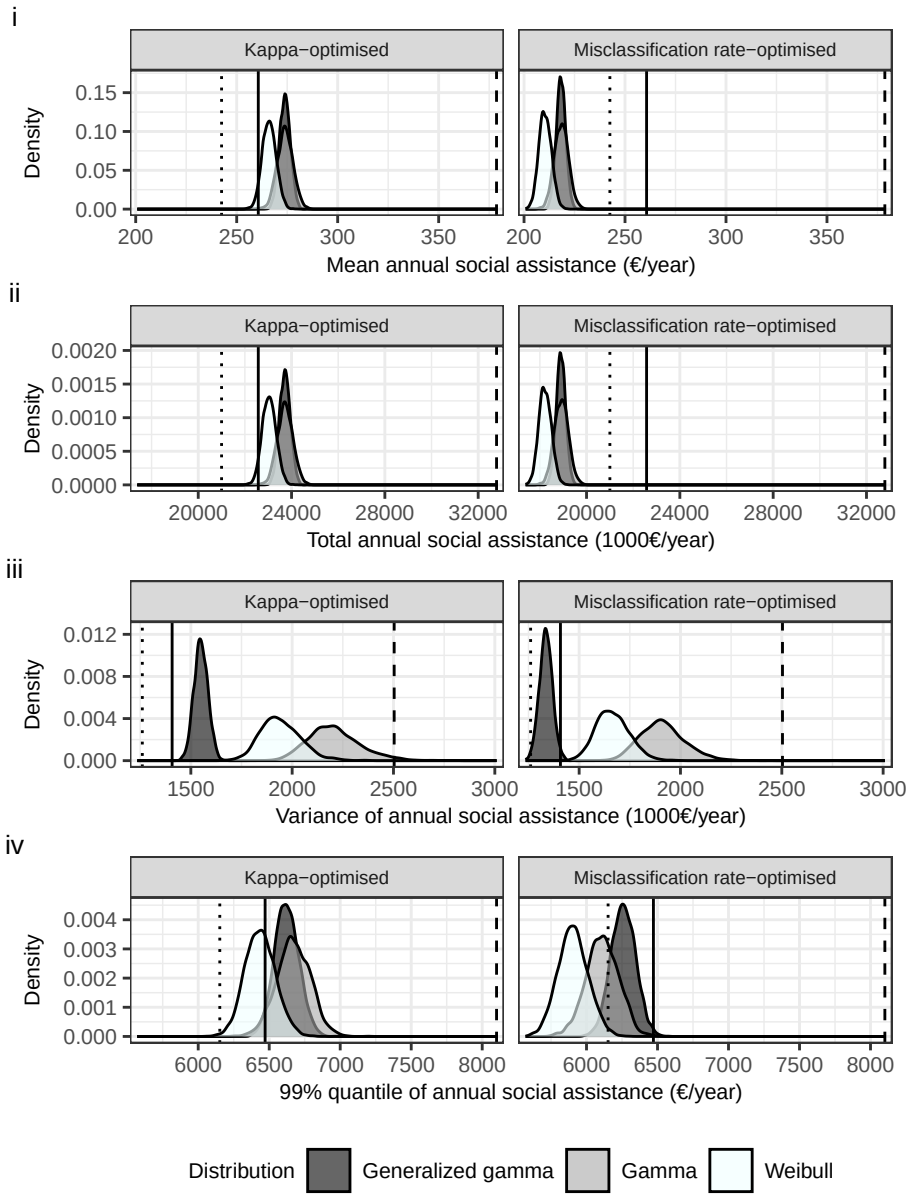

Figure 7: The distributions of predicted (i) mean, (i) total, (ii) variance and (iv) 99% quantile annual social assistance distributions for each of the two-part models according to the continuous distribution assumptions and the two classification methods using classification limits. The solid vertical line shows the observed statistic using the sum of basic, preventative and supplementary social assistance. The dashed vertical line shows corresponding predictions of the SISU model for annual basic social assistance, and the dotted vertical line depicts each statistic for the empirical annual basic social assistance. The estimates have been calculated using the test data set, and the two-part model estimates have been simulated using 3000 posterior samples.

classification limit which maximised Cohen's kappa, and the misclassification rate-optimised classifier was selected by minimising the misclassification rate. The model candidates were compared to each other and the SISU model using an external validation procedure.

The developed two-part models showed improvement in discrimination between social assistance recipients and non-recipients. The initially developed PP classifier did not manage to improve the classification accuracy compared to the SISU model. However, the estimates of the false positive rate, false negative rate, and misclassification rate for the KO classifier were lower than those of the SISU model. Further, as expected, the misclassification was the lowest when the classification was conducted using the MRO classifier. First, the KO and MRO classifiers produced more accurate results than the PP classifier. Second, the findings suggest that the classification result is sensitive to the choice of the classification limit. Third, a reduction in the false negative rate was only achieved with the KO classifier and this reduction was rather modest in comparison to the SISU model.

First, the KO and MRO classifiers performed better than the PP classifier because the PP classifier incorporates more uncertainty into the receipt status predictions than the KO and MRO classifiers. The PP classifier includes both the parameter uncertainty and the uncertainty over the receipt status given $\hat{\pi}_i^{(s)}$ in the predictions. Because $\hat{\pi}_i^{(s)}$ is an estimate of the expected value of the posterior predictive distribution, and the predicted class is determined by the fixed threshold, the uncertainty over the receipt status given $\hat{\pi}_i^{(s)}$ is reduced for the KO and MRO classifiers. This reduction in uncertainty together with the classification limit selection strategy likely produced the improved classification results.

Overall, the classification results were likely improved because the logistic regression of the binary component provides an opportunity to frame the problem of choosing the social assistance recipient status as a prediction problem. In this way, other covariates may be utilised in the prediction of the social assistance receipt status and uncertainty regarding take-up is incorporated into the parameter estimates. Then, whereas the SISU model assumes all potential social assistance recipients take up the benefit, the logistic regression model enables making an educated guess on the social assistance receipt status. It is likely that this slightly more sophisticated prediction method in part allowed improvements in the false positive rate.

Further, the developed two-part models showed improvement in the predictive accuracy of the annual social assistance in comparison to the SISU model. The RMSE estimates were lower than the RMSE estimates of the SISU model when assuming either Weibull or generalised gamma distributions, regardless of the classification method. The KS test d statistic estimates were lower than the SISU model KS test d statistic estimates when using the KO and PP classifiers,

regardless of the distribution assumption. Further, regardless of the base distribution or the classification method, the mean, total, variance and 99% quantile of the empirical annual social assistance distribution were estimated more accurately on average than the SISU model. The most accurate predictions of these quantities other than 99% quantile were produced when the classification was conducted by using the PP classifier.

The use of the generalised linear modelling was probably able to correct some systematic error in the false positive rate identified in Section 2.3.2. The SISU model's estimation method does not aim to estimate the probability of observing a receipt status. However, regression analysis is a supervised learning method, where the regression parameters describe the postulated relationship between the outcome and the covariates, and the regression parameters are estimated using a training data set. Therefore, by adding variables such as income, main activity and housing tenure, the regression parameters use information from the data set to reflect that while certain inactive groups might have low income, they typically do not receive social assistance.

In general, the classification and predictive accuracy of the developed models was likely improved because the developed two-part models use more information to make predictions than the SISU model. The SISU model uses fewer variables because it is limited to the use of variables relevant to the social assistance legislation. Then, to make predictions following the rules of the social assistance legislation using the annual level of data, the SISU model has to calculate monthly averages of income and social assistance to produce annual estimates of social assistance. This significantly aggregates the data, and together with the uncertainty introduced by missing wealth and expenditure variables, this introduces error to the point predictions of the annual social assistance. On the other hand, the generalised linear modelling approach enables the use of additional variables associated with social assistance in predictions of basic social assistance. Moreover, the statistical modelling approach does not have to make predictions based on aggregated data, because predictions are not based on the rules of the social assistance legislation.

Additionally, among the developed models, assuming a generalised gamma distribution for the continuous component generated the highest LPPD and the lowest RMSE in comparison to the Weibull and gamma distributions, regardless of the classification method. No difference between the developed models was found with regard to the KS test d statistic given that the classification method was fixed. This suggests that between the developed two-part models, the generalised gamma distribution produced the highest predictive accuracy. The better performance of the generalised gamma distribution likely results from the additional shape parameter of the distribution, which allows the distribution more flexibility to fit the shape of the empirical annual social assistance distribution.

The problems arising from the static and deterministic approach in microsimulation have also been addressed via statistical modelling in previous studies. For example, several studies have developed methods to investigate the labour participation choices as a response to legislation reforms (Harju et al., 2018; Carnicelli et al., 2020; Ollonqvist et al., 2021). These studies often formulate a utility function, which is aimed to reflect how an individual conducts decision-making on the outcome of interest, such as perceived utility derived from working as in Carnicelli et al. (2020). The maximum frequently represents an optimal choice, and often the interest is which combination of input variables maximises the utility function. Notably, a study by Harju et al. (2018) aimed to extend the SISU model by developing a multinomial logit model to predict the number of work hours at a given wage. The model acted as a part of a utility function that was used to represent the optimal amount of expenditure and leisure time. The legislative reforms altered the net income of the individual, and this in turn impacted the choice of work hours at a given wage that maximised the utility function. This approach is more flexible than the developed models in the present study because it provides a mechanism to investigate the impact of legislative reforms.

The results demonstrate the application of a two-part model on annual social assistance estimation and suggest that this approach may generate more accurate predictions of social assistance status and amount of annual social assistance. The current study adds to the previous research which demonstrates that a modelling approach which addresses uncertainty arising from take-up behaviour and limitations in the data set may improve microsimulation results. Moreover, the current model extends the design of the SISU model by incorporating predictions of supplementary and preventative social assistance in addition to basic social assistance. This further corroborates the benefits of utilising a statistical model, which may be used to generate predictions when the admission of a benefit is discretionary.

Some limitations hindering the improvements in predictive accuracy and applicability of the model arise from limitations of the data set, limited predictor selection strategy, violation of the modelling assumptions, simplicity of the classification methods, subjectivity in comparison strategy of the developed models and the SISU model, and the implicit assumption that the social assistance legislation is fixed.

Many of the same shortcomings in the register data set (see Section 2.4 for details) which introduce error into the SISU model predictions introduce error into the developed models. The modest improvements in the false negative rate likely reflect that the annual level of the data set hinders the detection of those receiving the benefit for a brief period, and the included covariates were unable to capture the features of this group.

Second, the modelling outcomes would have probably benefited from a more

extensive predictor selection strategy. In this study, only 13 alternatives of the predictor structure were considered in order to restrict computation time. A more extensive strategy would have been to conduct a more in-depth literature review on which factors might increase or decrease the probability of financial difficulties. On the other hand, some regularising prior distributions (e.g. Park and Casella, 2008) would have also provided an alternative strategy for predictor selection.

Third, the generalised regression models assume linearity and additivity of the linear predictor with respect to the quantity of the conditional distribution of the response variable after transformation with the link function, and it is reasonable to assume that this assumption might not hold for some selected predictors (such as salary, labour market subsidy or other social benefits) in predicting either the probability to observe social assistance receipt or the annual social assistance amount. Therefore, the model might be further improved by investigating and considering the non-linearity of the variables, such as by using splines (e.g. DiMatteo et al., 2001).

While the choice of classification limit showed improved discrimination of social assistance recipients and non-recipients, it introduced systematic error in the annual social assistance predictions. This effect was highlighted by findings that the PP classifier generated more accurate estimates of the mean, total and variance than the KO and MRO classifiers. The error likely results because the classification limit is selected after model fitting, that is, the classification limit is not an estimated parameter. Therefore, the development of the binary component could have benefited from the exploration of other classification methods, such as tree-based classifiers or neural networks (see e.g. Hastie et al., 2009).

Moreover, the comparison of estimates from the SISU model and the developed two-part model is hindered due to different modelling and estimation approaches. First, the SISU model and the developed model predict different response variables, which confuses comparisons of the calculated statistics. However, the preventative and supplementary assistance types form a very small proportion of total social assistance. More critically, the SISU model uses a deterministic estimation approach, and as a result, the uncertainty in the social assistance estimates is difficult to measure. In contrast, the developed model includes parameter uncertainty into its predictions and the predictive accuracy is measured through predictive performance on a test data set. In effect, the comparison strategy of the SISU model and the developed models is subjective in the sense that accuracy was chosen to be measured according to a certain collection of statistics and a selection of different statistics or another comparison strategy could have led to a different result.

Finally, the current model is developed by implicitly assuming that the 2019 social assistance legislation is in effect. The present model does not include a

mechanism to make modifications to the legislation and to investigate the effects of legislative reforms. This means that the process of how changes in legislation influence the eligibility of the house-dwelling unit, the amount of social assistance granted or the behaviour of the potential social assistance recipients is not defined. Therefore, the model is not suitable for investigation of the impact of legislation reforms. While this was not a modelling target, the lack of a mechanism to make modifications to the legislation hinders the application of the model.

The developed two-part models and the SISU model could be improved in the future in several ways. One evident opportunity would be to obtain a register data set recorded on a monthly level and to develop a model which predicts the amount of social assistance on a monthly basis. This type of modelling approach would most likely improve the false negative rate because fluctuations in income over the course of the year could be detected, such as a brief unemployment period. This approach would likely improve the SISU model's predictive accuracy even without the incorporation of statistical modelling components.

Moreover, modelling the development of social assistance over time would most likely further improve the estimation accuracy because it is reasonable to assume that the social assistance estimates in part depend on time. For example, the previous year's income could predict social assistance receipt in the following year. Estimates of the time dependency might also improve the estimates of social assistance expenditure in the future. Varying methods for time-dependent microsimulation of social benefits have already been applied to other topics in Finland, such as to social and health care client fees (Aaltonen et al., 2023) and the statutory pension system of Finland (Salonen, 2020; Salonen et al., 2019).

Further, in theory, a causal modelling approach (Pearl, 2009) could allow the estimation of causal effects of legislation reforms on an outcome of interest. First, devising a causal model would define the processes that should be modelled in order to estimate the causal effects of legislation reforms. Second, this approach would provide a natural framework to combine the rules of the legislation and uncertainties related to social assistance recipient status through structural equations. This would be significant because it's probable that the most successful microsimulation framework predicting social assistance would probably aim to combine aspects of the known data-generating process determined by the social assistance legislation and statistical modelling to account for uncertainties resulting from missing information and take-up behaviour.

In conclusion, the present study demonstrates that the application of a two-part model can improve both the classification accuracy of social assistance recipients and the predictive accuracy of annual social assistance estimates. The implication of the findings is to suggest how to incorporate uncertainty about the take-up of social assistance into predictions of annual social assistance. Despite the lack of

a mechanism to simulate the effects of legislation reforms, the present study recognizes a future research opportunity for causal modelling in microsimulation and the two-part model framework could potentially be extended with an appropriate causal model. Finally, this study contributes to the understanding of possible development opportunities of the SISU microsimulation model by suggesting how microsimulation methods could be augmented with statistical models.

References

Aaltonen, K., Tervola, J., and Heino, P. (2023). Analysing the effects of healthcare payment policies on poverty: a microsimulation study with real-world healthcare data. *International Journal of Microsimulation*, 16(1):89–107.

Ahola, E. and Hiilamo, H. (2013). Köyhyyttä helsingissä: Toimeentulotuen saajat ja käyttö 2008–2010. *Kansaneläkelaitos*.

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, 50(271):901–908.

Betancourt, M. (2016). Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1604.00695*.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Carnicelli, L., Kanninen, O., Karhunen, H., Kosonen, T., and Ravaska, T. (2020). Modeling family leave policies. *Prime Minister's Office*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cooper, N. J., Sutton, A. J., Mugford, M., and Abrams, K. R. (2003). Use of Bayesian Markov chain Monte Carlo methods to model cost-of-illness data. *Medical Decision Making*, 23(1):38–53.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071.

Fahrmeir, L., Tutz, G., Hennevogl, W., and Salem, E. (1994). *Multivariate statistical modelling based on generalized linear models*. Springer.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition.

Harju, J., Kyyrä, T., Kärkkäinen, O., Matikka, T., and Ojala, L. (2018). Työn tarjonnan mallintaminen osana talouspolitiikan vaikutusarviointia. *Valtioneuvoston kanslia*.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd edition.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hyndman, R. J. and Grunwald, G. K. (2000). Applications: generalized additive modelling of mixed distribution Markov models with application to Melbourne's

rainfall. *Australian & New Zealand Journal of Statistics*, 42(2):145–158.

Ilmakunnas, I., Kauppinen, T. M., and Kestilä, L. (2015). Sosioekonomisten syrjäytymisriskien kasautuminen vuonna 1977 syntyneillä nuorilla aikuisilla. *Yhteiskuntapolitiikka*, 80(3):247–262.

Ilmakunnas, I. and Moisio, P. (2019). Social assistance trajectories among young adults in Finland: What are the determinants of welfare dependency? *Social Policy & Administration*, 53(5):693–708.

Isotalo, E., Kyyrä, T., Lähdemäki, S., Pesola, H., Ravaska, T., Suhonen, T., and Villanen, J. (2022). Koronakriisin taloudellisten vaikutusten kohdentuminen. *Valtioneuvoston kanslia*.

Jauhiainen, S. and Korpela, T. (2019). Toimeentulotuen saajien elämäntilanne, asuminen ja työnteko. *Valtioneuvoston kanslia*.

Kauppinen, T. M., Angelin, A., Lorentzen, T., Bäckman, O., Salonen, T., Moisio, P., and Dahl, E. (2014). Social background and life-course risks as determinants of social assistance receipt among young adults in Sweden, Norway and Finland. *Journal of European Social Policy*, 24(3):273–288.

Kotamäki, M., Mattila, J., and Tervola, J. (2017). Turning static pessimism to dynamic optimism: An ex-ante evaluation of unemployment insurance reform in Finland. *The Social Insurace Institute of Finland*.

Kuivalainen, S. (2007). Toimeentulotuen alikäytön laajuus ja merkitys. *Yhteiskuntapolitiikka*, 72(1):49–56.

Liu, L., Shih, Y.-C. T., Strawderman, R. L., Zhang, D., Johnson, B. A., and Chai, H. (2019). Statistical analysis of zero-inflated nonnegative continuous data. *Statistical Science*, 34(2):253–279.

Livingstone, S., Betancourt, M., Byrne, S., and Girolami, M. (2019). On the geometric ergodicity of Hamiltonian Monte Carlo. *Bernoulli*, 25(4A):3109–3138.

Mesiäislehto, M., Elomäki, A., Närvi, J., Simanainen, M., Sutela, S., and Räsänen, T. (2022). The gendered impacts of the Covid-19 crisis in Finland and the effectiveness of the policy responses. *Finnish Institute for Health and Welfare*.

Metz, C. E. (1978). Basic principles of ROC analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier.

Moulton, L. H. and Halsey, N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, 51(4):1570–1578.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G. L., and Xiao-Li, M., editors, *Handbook of Markov Chain Monte Carlo*. CRC Press.

Neelon, B., O'Malley, A. J., and Smith, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research, Part 1: background and overview. *Statistics in Medicine*, 35(27):5070–5093.

Neelon, B., Zhu, L., and Neelon, S. E. B. (2015). Bayesian two-part spatial models for semicontinuous data with application to emergency department expenditures. *Biostatistics*, 16(3):465–479.

Ollonqvist, J., Tervola, J., Pirttilä, J., and Thoresen, T. O. (2021). The distributional effects of tax-benefit policies: A reduced form approach with application to Finland. *Finnish Institute for Health and Welfare*.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pearl, J. (2009). *Causality*. Cambridge University Press.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Salonen, J. (2020). *New methods in pension evaluation: Applications of trajectory analysis and dynamic microsimulation*. PhD thesis, Tampere University and Finnish Centre for Pensions.

Salonen, J., Tikanmäki, H., and Nummi, T. (2019). Using trajectory analysis to test and illustrate microsimulation outcomes. *International Journal of Microsimulation*, 12(2):3–17.

Schröer, G. (1991). *Computergestützte statistische Inferenz am Beispiel der Kolmogorov-Smirnov-Tests*. PhD thesis, Diplomarbeit Universität Osnabrück.

Schröer, G. and Trenkler, D. (1995). Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples. *Computational statistics & data analysis*, 20(2):185–202.

Sivula, T., Magnusson, M., Matamoros, A. A., and Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv preprint arXiv:2008.10296*.

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192.

Stacy, E. W. and Mihram, G. A. (1965). Parameter estimation for a generalized gamma distribution. *Technometrics*, 7(3):349–358.

Stan Development Team (2022). *Stan Modeling Language User's Guide and Reference Manual*.

Stan Development Team (2023). RStan: the R interface to Stan. R package version 2.32.3.

Statistics Finland (2018). *SISU–malli: Käyttöopas tulonsiirtojen ja verotuksen mikrosimulointiin*.

Tanhua, H. and Kiuru, S. (2020). Toimeentulotuki 2019: Toimeentulotuen saajien määrä laski. *Terveyden ja hyvinvoinnin laitos*.

Tervola, J., Mukkila, S., Ilmakunnas, I., Korpela, T., Mattila, H., Syväoja, J., Mäkinen, L., Okkonen, K.-M., Ollonqvist, J., Moisio, P., Keso, I., Liuttu, M., Mattila, J., Tuovinen, A.-K., and Martikainen, J. (2023). Perusturvan riittävyy-

den arviointiraportti 2019–2023. *Terveyden ja Hyvinvoinnin laitos*.

Vaalavuo, M. (2016). The development of healthcare use among a cohort of Finnish social assistance clients: testing the social selection hypothesis. *Sociology of Health & Illness*, 38(8):1272–1286.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian analysis*, 16(2):667–718.

Warrens, M. J. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of classification*, 25:195–208.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Appendix

Project code and results which met the data privacy guidelines are available in the GitHub repository: `https://github.com/aeritala/sisu_social_assistance`

Table A1: Mean, first quantile (Q1), third quantile (Q3) of annual social assistance, and proportion of social assistance recipients in groups defined by attributes of the house-dwelling units. Housing allowance refers to general housing allowance, housing allowance for students and housing allowance for pensioners.

Variable	Mean (€/year)	Q1 (€/year)	Q3 (€/year)	Prop. of SA recipients (%)
Sex of the reference person				
Male	2832.96	774.51	4047.34	4.82
Female	2815.68	718.78	3987.60	4.28
Age of the reference person				
[0, 18)	2169.63	671.02	3054.60	0.14
[18, 25)	2995.57	906.67	4465.96	1.83
[25, 35)	3144.34	943.50	4458.05	2.21
[35, 45)	3182.77	866.62	4524.16	2.33
[45, 65)	2427.83	630.00	3386.02	1.97
[>65]	1263.28	278.56	1567.47	0.63
House-dwelling unit structure				
Other	2571.97	598.37	3871.60	0.55
Single	2633.55	741.79	3716.76	5.21
Single with children	3173.34	901.83	4537.75	1.32
Couple with children	2560.05	623.05	3591.14	0.87
Couple	3635.33	801.70	5465.67	0.98
Unknown	3478.95	757.97	4868.68	0.17
Housing tenure				
Other	2602.15	616.31	3763.93	0.74
Owner-occupied	2084.28	474.00	2707.35	1.10
Part-ownership	2600.80	723.89	3648.61	0.10
Rented-dwelling	2964.37	832.89	4201.45	7.17
Main activity of the reference person on the last day of the year				
Employed	1889.97	531.38	2539.40	2.24
Long-term unemployed	3694.95	1399.61	5239.03	2.57
Student	3187.32	1105.21	4555.64	1.20
On pension	1431.33	299.00	1822.66	1.58
Entrepreneur	2362.45	610.50	3290.54	0.13
Other	4041.85	1636.70	5747.45	1.39
Education level of the reference person				
Comprehensive	3420.44	955.56	5057.65	3.75
Upper secondary	2420.08	660.81	3365.89	4.33
Post-secondary non-tertiary	2058.61	589.44	2909.89	0.04
Lowest tertiary	2295.71	549.85	3103.50	0.30
Bachelor's or equiv.	2420.50	650.46	3332.04	0.45
Masters or equiv. and second stage of tertiary	2366.64	647.66	3229.01	0.24
Immigration status of the reference person				
No	2629.28	685.80	3722.69	7.40
Yes	3670.60	1211.63	5203.87	1.71

Table A1: *(continued)*

Variable	Mean (€/year)	Q1 (€/year)	Q3 (€/year)	Prop. of SA recipients (%)
Total salary per house-dwelling unit (1000€/year)				
[0, 0.5k)	3273.77	941.63	4760.47	5.26
[0.5k, 3k)	3270.56	1108.10	4611.84	0.72
[3k, 7k)	2688.03	804.43	3792.64	0.65
[7k, 10k)	2154.90	669.39	2822.60	0.38
[10k, 20k)	1861.91	559.49	2469.28	0.81
[>20k]	1613.23	421.93	2086.20	1.29
Total housing allowance per house-dwelling unit (€/year)				
[0, 0.5k)	1541.33	369.70	1989.16	1.30
[0.5k, 1k)	1678.05	442.62	2131.02	0.39
[1k, 3k)	2124.63	588.64	2930.34	2.03
[3k, 5k)	3197.24	1135.71	4519.91	3.88
[>5k]	4215.11	1244.74	6153.20	1.50
Total labour market subsidy per house-dwelling unit (number of days)				
[<30)	2582.97	540.62	3646.05	5.00
[30, 60)	3267.84	875.43	5030.48	0.31
[60, 100)	3111.29	938.60	4529.16	0.37
[100, 200)	3011.61	1050.54	4263.81	0.92
[200, 300)	2889.39	1308.01	3824.67	2.18
[>300]	4864.85	2357.75	6674.96	0.32

Table A2: Descriptions of the covariates used in the best-performing logistic, gamma, Weibull and generalized gamma regression models.

Variable	Description	Unit
Intercept	Intercept, reference classes marked with (ref.)	
Age	Age of the reference person	under 18, 18–24 (ref.), 25–34, 35–44, 45–64, 65 or older
Sex	Sex of the reference person	male (ref.), female
House-dwel. structure	House-dwelling unit structure	single (ref.), single with children, couple, couple with children, other, unknown
Education level	Education level of the reference person	comprehensive school (ref.), upper secondary school, post-secondary non-tertiary education, lowest tertiary education, bachelor's or equivalent, master's or equivalent and second stage of tertiary education
Housing tenure	Housing tenure	rented (ref.), owner-occupied, part-ownership, other
Immigrant status	Immigrant status	no (ref.), yes
Municip. class	Municipality class	Helsinki (ref.), other Helsinki metropolitan area, middle-sized cities, other municipalities

Table A2: *(continued)*

Variable	Description	Unit
Communal house-dwel.	Communal house-dwelling unit status, indicator whether persons aged 18 or over that are not the spouse of the reference person live in the house-dwelling unit	no (ref.), yes
Unemp. duration	Unemployment duration of the reference person	number of months, 0–12
Income decile	Income decile	ordinal number, 0–9
Capital income	Capital income	€/year, total across all recipients in the house-dwelling unit
Labour market subsidy	Duration of labour market subsidy receipt	number of days, total across all recipients in the house-dwelling unit
Basic unemp.	Duration of basic unemployment allowance receipt	number of days, total across all recipients in the house-dwelling unit
Earnings-rel. unemp.	Duration of earnings-related unemployment allowance receipt	number of days, total across all recipients in the house-dwelling unit
Student aid	Duration of reference person's student aid receipt	number of months, 0–12
SISU soc.assist. cont.	Amount of annual basic social assistance predicted by the SISU model	€/year
SISU soc.assist. categ.	Categorical amount of annual social assistance predicted by the SISU model	0–580 €/year, 581–2424 €/year, 2425–5482 €/year, more than 5483 €/year
Housing allowance	Amount of general housing allowance, housing supplement for students and housing allowance for pensioners	€/year, total across all recipients in the house-dwelling unit
Family's benefits	Child benefit, maternity grant and child maintenance allowance	€/year, total across all recipients in the house-dwelling unit
Other social benefits	Total of other social benefits	€/year, total across all recipients in the house-dwelling unit
Pension	Amount of national pension, survivor's pension, inc. guarantee pension	€/year, total across all recipients in the house-dwelling unit
Main activity	Main activity of the reference person over the course of the year	employed (ref.), entrepreneur, on pension, student, other
Salary	Amount of annual salary	1000€/year divided by the number of adults (18-year-olds or older) in the house-dwelling unit

Table A3: Posterior means, standard errors, $\hat{R}$-values, bulk-ESS and tail-ESS values of the regression coefficients from the logistic regression model which was developed as the binary component of the two-part model. Estimates are on the scale of the linear predictor (logit). Detailed covariate descriptions are provided in the Appendix A2.

Variable	Levels	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
Intercept	Intercept	-13.40	1.69	1.00	5013.90	1602.78
Age	Under 18	-1.07	0.06	1.00	4404.69	2101.25
	25–34	-0.01	0.03	1.00	2931.03	2329.16

Table A3: *(continued)*

Variable	Levels	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
	35–44	-0.07	0.03	1.00	2836.63	2482.45
	45–65	-0.43	0.03	1.00	2296.36	2140.99
	65 or older	-1.88	0.04	1.00	3410.54	2637.57
Sex	Female	0.03	0.02	1.00	6645.70	2102.75
House-dwel. structure	Couple with children	0.21	0.03	1.00	3860.70	2448.68
	Couple	-0.57	0.03	1.00	3966.59	2421.87
	Unknown	-0.14	0.06	1.00	2356.84	2369.38
	Other	0.32	0.07	1.00	4276.69	2082.13
	Single with children	0.75	0.03	1.00	3538.05	2299.78
Main activity	On pension	-0.40	0.04	1.00	2692.28	2335.85
	Other	0.30	0.03	1.00	3322.49	2477.72
	Student	0.37	0.06	1.00	3026.40	2229.02
	Entrepreneur	-1.50	0.06	1.00	3726.63	2604.11
Housing tenure	Owner-occupied	-0.98	0.05	1.00	2268.32	2340.75
	Part-ownership	-0.06	0.08	1.00	3174.80	2144.71
	Rented dwelling	0.78	0.05	1.00	2199.02	1957.56
Communal house-dwel.	Yes	0.69	0.03	1.00	3764.84	2328.97
Education level	Upper second.	-0.51	0.02	1.00	4883.87	2587.35
	Post-second. non-tert.	-0.47	0.10	1.00	6097.94	1731.80
	Lowest tertiary	-0.85	0.04	1.00	5024.80	2506.02
	Bachelor's or equiv.	-1.17	0.03	1.00	4867.17	2286.11
	Master's or equiv. and second stage of tert.	-1.54	0.04	1.00	5723.35	2371.87
Municip. class	Other Helsinki metrop. area	0.13	0.03	1.00	3685.76	2531.76
	Middle-sized cities	-0.05	0.02	1.00	3379.69	2169.19
	Other municipalities	-0.03	0.03	1.00	3558.80	2483.53
Immigrant status	Yes	-0.03	0.02	1.00	5815.45	2323.49
Unemp. duration		0.55	0.01	1.00	5007.49	2390.56
Income decile		0.04	0.05	1.00	4607.36	2632.03
Salary		-40.13	0.73	1.00	2694.30	2318.89
Capital income		-0.04	1.98	1.00	6098.65	2300.92
Housing allowance		27.85	1.99	1.01	6231.75	1906.15
Labour market subsidy		15.82	0.37	1.00	4818.50	2256.25
Basic unemp.		1.25	0.09	1.00	6189.29	2443.06
Earnings-rel. unemp.		-5.92	0.30	1.00	5194.60	2274.88
Pension		3.80	2.08	1.00	7928.79	1744.57
Student aid		-0.28	0.02	1.00	3604.10	2403.23
SISU soc.assist. cont.		3.68	1.95	1.00	6596.88	2034.56
Family's benefits		5.54	1.96	1.00	4947.15	2114.59
Other social benefits		3.31	1.91	1.00	6509.22	2199.45

Table A4: Posterior means, standard errors, $\hat{R}$-values, bulk-ESS and tail-ESS values of the regression coefficients from the gamma regression model which was developed as the continuous component of the two-part model. Estimates are on the scale of the linear predictor (log). The star (*) denotes an interaction term. Detailed covariate descriptions are provided in the Appendix A2.

Variable	Levels	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
Intercept	Intercept	-5.20	1.18	1.00	3400.76	2269.80
Age	Under 18	-0.32	0.04	1.00	3973.57	2773.50
	25–34	0.03	0.02	1.00	1841.83	2060.45
	35–44	0.10	0.02	1.00	1638.78	2107.88
	45–65	0.02	0.02	1.00	1449.68	1717.61
	65 or older	-0.13	0.03	1.00	1879.15	1996.08
Sex	Female	0.01	0.01	1.00	4291.29	2083.62
House-dwel. structure	Couple with children	-4.75	0.54	1.00	1490.06	1890.81
	Couple	-3.49	0.59	1.00	1513.56	1958.37
	Unknown	-0.03	0.81	1.00	1976.84	2023.77
	Other	-2.58	0.82	1.00	1710.46	1898.43
	Single with children	2.27	0.52	1.00	1478.45	1743.85
Main activity	On pension	-0.52	0.02	1.00	1479.47	1945.06
	Other	0.16	0.02	1.00	2184.89	2202.72
	Student	0.13	0.04	1.00	2012.60	2278.43
	Entrepreneur	-0.39	0.04	1.00	3584.36	2691.67
Housing tenure	Owner-occupied	-0.21	0.04	1.00	1847.44	2182.48
	Part-ownership	-0.19	0.06	1.00	2604.49	2394.45
	Rented dwelling	-0.19	0.03	1.00	1771.23	2132.80
Communal house-dwel.	Yes	0.20	0.02	1.00	4732.66	2253.10
Education level	Upper second.	-0.14	0.01	1.00	4047.97	2066.74
	Post-second. non-tert.	-0.14	0.07	1.00	3588.72	2178.49
	Lowest tertiary	-0.18	0.03	1.00	4273.60	2244.05
	Bachelor's or equiv.	-0.21	0.02	1.00	3556.27	2367.53
	Master's or equiv. and second stage of tert.	-0.24	0.03	1.00	4025.04	2440.65
Municip. class	Other Helsinki metrop. area	0.01	0.02	1.00	2402.82	2482.91
	Middle-sized cities	-0.07	0.01	1.00	1865.17	2157.85
	Other municipalities	-0.09	0.02	1.00	1942.73	2299.13
Immigrant status	Yes	0.02	0.01	1.00	3641.73	2272.71
Unemp. duration		0.10	0.01	1.00	4452.32	2028.62
Income decile		2.01	0.04	1.00	2451.00	2052.00
Salary		-20.61	0.55	1.00	1401.14	1949.38
Capital income		0.01	1.48	1.00	4731.30	2288.93
Housing allowance		5.76	1.44	1.00	4318.82	2293.19
Labour market subsidy		-2.10	0.20	1.00	2516.25	2529.34
Basic unemp.		-0.48	0.05	1.00	3787.50	2571.50
Earnings-rel. unemp.		-3.45	0.19	1.00	2948.10	2406.66
Pension		-1.50	1.50	1.00	4485.92	2401.68

Table A4: *(continued)*

Variable	Levels	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
Student aid		0.03	0.02	1.00	2020.38	2154.58
SISU soc.assist. categ.	581–2424	0.42	0.02	1.00	2990.11	1946.91
	2425—5482	0.71	0.02	1.00	2694.52	2483.96
	more than 5483	1.16	0.02	1.00	2457.33	2668.87
Family's benefits		1.58	1.49	1.00	4747.36	2146.49
Other social benefits		-1.04	1.50	1.00	4015.95	2393.73
House-dwel. structure * Salary	Salary * Couple with children	-6.27	0.65	1.00	1490.71	1988.56
	Salary * Couple	-4.39	0.72	1.00	1515.35	1854.35
	Salary * Unknown	0.11	0.97	1.00	1998.11	1937.93
	Salary * Other	-3.42	1.00	1.00	1702.53	2004.18
	Salary * Single with children	2.53	0.63	1.00	1485.13	1704.32

Table A5: Posterior mean, standard error, $\hat{R}$-value, bulk-ESS and tail-ESS value of the shape parameter from the gamma regression model which was developed as the continuous component of the two-part model.

	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
Shape	1.48	0.01	1.00	5940.03	2251.1

Table A6: Posterior means, standard errors, $\hat{R}$-values, bulk-ESS and tail-ESS values of the regression coefficients from the Weibull regression model which was developed as the continuous component of the two-part model. Estimates are on the scale of the linear predictor (log). The star (*) denotes an interaction term. Detailed covariate descriptions are provided in the Appendix A2.

Variable	Levels	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
Intercept	Intercept	-4.77	1.18	1.00	3408.13	2178.74
Age	Under 18	-0.29	0.04	1.00	3874.44	2137.19
	25–34	0.03	0.01	1.00	2261.70	2488.46
	35–44	0.10	0.01	1.00	1873.57	2249.20
	45–65	0.02	0.02	1.00	1934.71	2168.79
	65 or older	-0.13	0.02	1.00	2380.34	1968.65
Sex	Female	0.01	0.01	1.00	5183.29	2305.77
House-dwel. structure	Couple with children	-4.66	0.52	1.00	1942.84	2068.14
	Couple	-3.52	0.54	1.00	1873.02	1955.85
	Unknown	-0.10	0.81	1.00	2075.88	2023.63
	Other	-2.80	0.81	1.00	2279.45	1982.32
	Single with children	2.65	0.49	1.00	1741.32	1885.72
Main activity	On pension	-0.49	0.02	1.00	1622.38	2158.82
	Other	0.15	0.02	1.00	1984.51	2206.18
	Student	0.12	0.04	1.00	2357.09	2286.07

Table A6: *(continued)*

Variable	Levels	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
	Entrepreneur	-0.36	0.04	1.00	3962.35	1889.71
Housing tenure	Owner-occupied	-0.20	0.03	1.00	1801.67	2098.76
	Part-ownership	-0.20	0.05	1.00	2244.41	2378.89
	Rented dwelling	-0.20	0.03	1.00	1875.36	2112.55
Communal house-dwel.	Yes	0.21	0.02	1.00	4476.24	2237.82
Education level	Upper second.	-0.14	0.01	1.00	3524.90	2333.70
	Post-second. non-tert.	-0.14	0.06	1.00	4491.32	2455.32
	Lowest tertiary	-0.17	0.03	1.00	5226.77	2400.80
	Bachelor's or equiv.	-0.20	0.02	1.00	4041.12	2335.76
	Master's or equiv. and second stage of tert.	-0.22	0.03	1.00	4424.38	2405.83
Municip. class	Other Helsinki metrop. area	0.01	0.02	1.00	3227.36	2180.94
	Middle-sized cities	-0.07	0.01	1.00	2439.72	2233.61
	Other municipalities	-0.09	0.01	1.00	2554.79	2224.22
Immigrant status	Yes	0.03	0.01	1.00	3797.32	2005.88
Unemp. duration		0.10	0.01	1.00	4807.74	2091.34
Income decile		2.00	0.04	1.00	3013.12	2137.58
Salary		-20.28	0.50	1.00	1609.68	2012.79
Capital income		-0.01	1.46	1.00	5674.43	2342.41
Housing allowance		6.14	1.47	1.00	3915.08	1950.61
Labour market subsidy		-2.27	0.18	1.00	2831.69	2208.55
Basic unemp.		-0.47	0.04	1.00	4171.21	2348.21
Earnings-rel. unemp.		-3.27	0.17	1.00	4227.26	2276.96
Pension		-1.62	1.50	1.00	4675.79	1928.38
Student aid		0.03	0.01	1.00	2638.27	2570.37
SISU soc.assist. categ.	581–2424	0.40	0.01	1.00	2900.40	2143.62
	2425—5482	0.68	0.01	1.00	3090.71	2390.19
	more than 5483	1.13	0.02	1.00	2404.36	2373.86
Family's benefits		1.79	1.49	1.00	4516.02	2213.68
Other social benefits		-1.20	1.49	1.00	6068.92	2102.18
House-dwel. structure * Salary	Salary * Couple with children	-6.18	0.63	1.00	1952.02	2010.10
	Salary * Couple	-4.45	0.66	1.00	1885.01	1877.21
	Salary * Unknown	0.01	0.96	1.00	2082.15	2055.54
	Salary * Other	-3.71	0.98	1.00	2266.37	1945.65
	Salary * Single with children	2.97	0.60	1.00	1741.42	1902.89

Table A7: Posterior mean, standard error, $\hat{R}$-value, bulk-ESS and tail-ESS value of the shape parameter from the Weibull regression model which was developed as the continuous component of the two-part model.

	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
Shape	1.31	0.01	1.00	7109.39	2060.56

Table A8: Posterior means, standard errors, $\hat{R}$-values, bulk-ESS and tail-ESS values of the regression coefficients from the generalized gamma regression model which was developed as the continuous component of the two-part model. Estimates are on the scale of the linear predictor (log). Detailed covariate descriptions are provided in the Appendix A2.

Variable	Levels	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
Intercept	Intercept	0.40	0.04	1.00	1651.40	1896.91
Age	Under 18	-0.33	0.03	1.00	3783.55	2373.78
	25–34	0.02	0.01	1.00	2826.99	2427.10
	35–44	0.05	0.01	1.00	2763.67	2132.83
	45–65	-0.04	0.01	1.00	2802.55	2503.82
	65 or older	-0.16	0.02	1.00	3262.13	2340.47
Sex	Female	-0.01	0.01	1.00	4937.26	2139.03
House-dwel. structure	Couple with children	0.12	0.02	1.00	1815.83	1865.67
	Couple	-0.13	0.03	1.00	1856.12	2310.33
	Unknown	0.02	0.05	1.00	2403.41	2242.21
	Other	-0.04	0.04	1.00	2809.06	2027.21
	Single with children	0.04	0.02	1.00	2320.20	2473.19
Main activity	On pension	-0.42	0.02	1.00	2612.10	2402.47
	Other	0.13	0.01	1.00	3078.20	2302.36
	Student	0.13	0.03	1.00	2734.84	2398.42
	Entrepreneur	-0.49	0.04	1.00	3650.46	2269.49
Housing tenure	Owner-occupied	-0.10	0.03	1.00	1925.05	2114.43
	Part-ownership	-0.10	0.04	1.00	3101.16	2119.54
	Rented dwelling	-0.10	0.02	1.00	1999.26	1997.60
Communal house-dwel.	Yes	0.12	0.01	1.00	4085.70	2175.41
Education level	Upper second.	-0.14	0.01	1.00	3414.88	2267.64
	Post-second. non-tert.	-0.16	0.05	1.00	4052.05	1948.48
	Lowest tertiary	-0.16	0.02	1.00	4363.56	2094.91
	Bachelor's or equiv.	-0.20	0.02	1.00	4259.62	2339.45
	Master's or equiv. and second stage of tert.	-0.29	0.02	1.00	4163.95	1906.23
Municip. class	Other Helsinki metrop. area	0.03	0.01	1.00	3392.78	2448.50
	Middle-sized cities	0.03	0.01	1.00	2860.67	2003.48
	Other municipalities	0.02	0.01	1.00	2940.09	2698.50
Immigrant status	Yes	-0.03	0.01	1.00	4994.11	2037.95
Unemp. duration		0.04	0.00	1.00	4544.57	2246.59
Income decile		0.51	0.01	1.00	2826.64	2314.94
Salary		-0.75	0.03	1.00	2042.78	1889.34
Capital income		-0.66	0.10	1.00	4009.95	1855.39

Table A8: *(continued)*

Variable	Levels	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
Housing allowance		0.11	0.00	1.00	3981.87	2420.10
Labour market subsidy		-0.06	0.00	1.00	3009.46	2489.20
Basic unemp.		-0.03	0.00	1.00	4037.55	2444.62
Earnings-rel. unemp.		-0.08	0.00	1.00	4264.17	2800.45
Pension		-0.08	0.00	1.00	4433.19	2714.36
Student aid		-0.04	0.01	1.00	3160.13	1787.44
SISU soc.assist. categ.	581–2424	0.30	0.01	1.00	3029.98	2509.02
	2425—5482	0.54	0.01	1.00	2364.54	2155.07
	more than 5483	0.80	0.01	1.00	1930.49	2077.36
Family's benefits		-0.02	0.00	1.00	4404.61	2540.67
Other social benefits		-0.06	0.00	1.00	4345.70	2653.19
House-dwel. structure * Salary	Salary * Couple with children	-0.32	0.03	1.00	1715.88	1936.55
	Salary * Couple	-0.23	0.04	1.00	1902.21	1686.70
	Salary * Unknown	0.14	0.06	1.00	3009.60	2350.37
	Salary * Other	-0.35	0.07	1.00	3229.23	1719.36
	Salary * Single with children	0.07	0.03	1.00	2553.50	2720.63

Table A9: Posterior mean, standard error, $\hat{R}$-value, bulk-ESS and tail-ESS value of the scale (β) and shape (δ) parameters from the generalized gamma regression model which was developed as the continuous component of the two-part model.

	Estimate	SE	$\hat{R}$	Bulk-ESS	Tail-ESS
β	1137.23	40.16	1.00	1545.69	1666.25
δ	0.92	0.01	1.00	1494.51	1669.25