Stockholms universitet

Master's Thesis

Department of Statistics

No. 2011:1

Small Area Estimation of Rare Events: Estimating Victimization Rates in the Swedish Crime Survey

Måns Magnusson

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Acknowledgment

This report/thesis has been written and produced both as a report on statistical and methodological questions at the National Council for Crime Prevention (NCCP) and as a master thesis at the Department of Statistics, Stockholm University. For this reason the same report will be published both as a methodological paper at the NCCP and as a master thesis at Stockholm University. I, Måns Magnusson, am the author of the publications that are identical (except for the covers).

Dan Hedlin from Statistics Sweden and the Department of Statistics, Stockholm University, has been the supervisor of the study and I would like to thank him for all the help and guidance in the production of this report/thesis.

Abstract

In this report different design-based estimators are studied as to how well they perform when it comes to estimating victimization rates for different criminal offences at municipality and county level domains in the Swedish Crime Survey.

Linear and logistic generalized regression estimators with different auxiliary information, such as domain specific intercepts, reported number of crimes at the municipality level and reported crimes at the respondent level (i.e. whether the respondent reported the crime or not) are studied. Earlier research suggests that the classical Wald interval works poorly in the situation of rare events and small samples. The Wilson interval has been suggested as an alternative.

The estimators are examined through a design-based simulation study based on real life data from the Swedish Crime Survey.

One major finding of the study is that using reported crime at the individual level (if the person has reported an offence or not) reduces the MSE of the GREG estimators by 10-20 percent. The smaller the dark figure, the better the auxiliary variable is to explain the variance and hence the more precise the estimator. To get the same effect by increasing sample sizes one would have to increase the sample size by as much as 25 %. The other auxiliary variables, such as domain specific intercepts and the number of reported crimes at municipality level, are not as good.

Another finding is that there is no large gain by using the logistic GREG instead of the ordinary linear GREG. Previous research suggests that the logistic GREG should outperform the linear GREG, if the auxiliary information is strong enough. In this study the auxiliary information was not strong enough to get any larger gains.

A third finding is that the Wald interval performs badly, even for quite large sample sizes, such as $101 < n < 349$. The Wilson interval, adapted for the GREG estimator, proved to work much better. For less rare offences such as "Crime against persons" the Wilson interval works for all domain sizes. For the less common crime types, such as robbery and sexual assault, the Wilson interval works well for sample sizes with $n > 100$.

Based on the results of the study one main conclusion, all crime types can be estimated down to $n > 100$ and for some, more common crime types such as harassment, estimates can be produced for sample sizes $n > 40$. For smaller sample sizes it is possible to produce estimates for "Crimes against persons", but the issue is if the intervals may be too wide to be of any practical use.

The possibility to use police records in the estimation phase should be studied further since this variable seems to be a strong auxiliary variable in estimation of crime rates.

Keywords:
Small area estimation, victimization, Swedish Crime Survey

Contents

1. Introduction

1.1. Background

In Sweden as well as in other countries there is an increased demand for statistical estimates of smaller and smaller geographical and demographical domains. These estimates are often needed for public policy and decision-making in different agencies and organizations. In register-based surveys or censuses, estimates for small domains are not a difficult problem since all elements are (in theory) observed and the sampling error is non-existent. In sample-based surveys, on the other hand, estimates for small domains become a problem when estimates are sought for smaller domains than the survey was originally intended for.

The problem for sample-based surveys will increase further if the estimates that are sought for are rare in the population as a whole. In these situations even quite large domains will end up with very few or even no elements with the observed characteristics. In the design-based framework the variance estimate will end up with zero if the Horvitz Thompson variance estimator is used and then an interval estimate will not be possible to produce. The smaller the domains get, the larger the problem will be. In really small areas there will probably be no objects in the domain with the rare characteristic.

Sweden's largest victimization survey, The Swedish Crime Survey (SCS), has a sample size of 20 000 each year, but when it comes to estimates at lower regional levels as counties ("län" in Swedish) and municipalities ("kommuner" in Swedish) the sample sizes get smaller. In some counties the sample size is just 500 and in the municipalities sample sizes can be as small as just a handful of respondents.

One of the major purposes of the SCS is to estimate the victimization rates of different offences and many of those offences are very rare. For instance, in the year of 2008, only 1 percent of the respondents reported that they had experienced robbery and only 0.7 percent of the Swedish households reported experience of grand theft auto (Irlander and Westfelt 2010:1, pp. 28, 38.). These low rates mean that there are a large number of small domains that will lack respondents having experienced robbery – even if the sample is large enough for other types of estimates.

One of the primary users of the SCS is the Swedish police. The police use the Swedish crime survey to benchmark between counties and compare the development of victimization rates over time in each county.

The only estimates published today regarding victimization are pooled estimates including a large number of different crimes called "Crimes against persons" and "Crimes against households", but this pooling is not of interest for all the main users. They are interested in each crime type such as robbery, assault etc.

To be able to compare changes in time and between counties the interval estimates should be as narrow as possible, both when it comes to the more rare types of crimes as well as the more common types.

1.2. Purpose of study

The purpose of this study is to examine different small area estimators and to examine how they perform in the presence of rare events. These estimators will be studied in the context victimization rates for different offences (some more rare than others) in the SCS. Statistics of interest are totals and proportions in different geographical domains such as counties (NUTS 3-level) and municipalities (NUTS 5-level). Estimates for different types of experienced offences are of interest, such as assault, robbery, burglary, threats, harassment and sexual offences. The aim of this study is also to be able to evaluate different estimators and examine if these estimators can be used in the production of official statistics.

1.3. Previous research and theory

1.3.1. Small area estimation - theory

The most widely used definition of small area estimation is given by Rao (2003) at the very first page: "Small area estimation is defined as estimates for domains that are too small, with regard to sample size, to support direct estimates of adequate precision." Direct estimates are defined as estimates based only on the sample within the specific domain. In Lehtonen and Pahkinen (2004) the definition of small area estimation is plainly "Estimates for domains with small sample sizes."

Both in ordinary sampling theory as well as in the area of small area estimation, there are three main schools of inference, design-based inference, model-based inference and Bayesian inference. A comparison between the different approaches to inference is given by Thorburn (2009):

Table 1: Different approaches to survey sampling (Thorburn 2009)

	Design-based	**Model-based**	**Bayesian**
Randomness/ Uncertainty	Home-made	Given by nature/ Frequency-based	Subjective/Rationality axioms
Main focus	Population	Parameters	Population
Parameters	Population values	Unknown/ Unobservable	Do not exist, but useful/ DeFinetti Theorem
Inference	Frequency-based	Frequency-based	Probability-based
Output	Point-estimates/Intervals	Point-estimates/ Confidence intervals	Posterior distributions/ means and variances
Possible Use	Not my problem!	Not my problem!	Rational interface with decisions

In Thorburn (2009) p. 14-16 three different approaches to inference is illustrated in both small area estimation and survey sampling. As presented in the table, the randomness and uncertainty are very different in the different schools. In the design-based approach there is randomness created by the sampling process whereas the population is treated as fixed. The model-based inference, on the other hand, assumes that a stochastic super population model creates the population. The Bayesians do not assume a population model or introduce their own, homemade, randomness. Instead the Bayesians take what they know into account and describe it as a probability distribution. The randomness in the Bayesian sense is more of an uncertainty than randomness (Thorburn 2009, pp. 14-16).

The main focus of the three approaches differs as well. The Bayesian approach and the design-based approach focus on the population values in the finite population. The model-based approach, on the other hand, sees the population as a sample of the super population and puts its focus on the parameters of the stochastic super population model (Thorburn 2009, pp. 14-16 and Ott 2007).

In the Bayesian approach the prior distribution, based on prior knowledge, is combined with the observed data to a posterior distribution. The posterior distributions are then used for inference to the population. This makes the Bayesian approach different from the other two schools. In both the model-based and the design-based schools of thought the inference is frequency-based. An interval estimate of 95% means that the calculated interval will cover the true value in the population (design-based) or the super population model parameters (model-based) in 95% of repeated samples (Thorburn 2009, pp. 14-16 and Ott 2007).

The last main difference between the different approaches to inference from sample surveys is that the end result of the design-based and the model-based approaches is only one estimate. When estimates have been calculated and disseminated the statistician's job is done. In the Bayesian approach the purpose of the posterior distribution is not only to disseminate statistics but also to be an aid in decision making. By using the posterior distribution, decision problems can be examined more in detail (Thorburn 2009, pp. 14-16).

In small area estimation all approaches are being used and studied (se for example Rao 2003). Since

the scientific area of small area estimation is, by definition, the problem of estimation areas that has been too small for design-based estimation, much focus has been put on using model-based and Bayesian methods. But also different design-based estimators have been proposed as solving some small area problems.

The second and third purposes of this study are to examine different estimators for small domains in the SCS that can be included in the production of official statistics. In official statistics today almost all estimators used are design-based and one of the main purposes of this study is to look into estimators that can be used for official statistics. For this reason as well as to limit the scope of this study only design-based estimators will be studied.

1.3.2. Design-based estimation

The basic philosophy in the design-based approach to survey sampling has been explained in part 1.2.1 above. The main focus of the design-based estimation is to estimate population parameters (as totals) in a finite population by "controlling" the uncertainty. A short introduction to notations and definitions is needed. The definitions and notations largely follow the definitions and notations of Särndal, Swensson and Wretman (1992) and Lehtonen and Pahkinen (2004).

A population, denoted U, contains N elements. A sample, denoted s, of (1 , ... , n) elements is chosen at random from the population. The probability of drawing a particular sample, s, is denoted p(s) and is called the *sampling design* where the space of all possible samples is denoted ζ. This sampling design can also be defined as the random variable S with the pdf p(s). Based on the sampling design the first order inclusion probability of element k, π_k, and second order inclusion probability of element k and l, π_{kl} can be calculated in the following way (Särndal, Swensson and Wretman 1992, pp. 24-33.):

$$\pi_k = \sum_{k \in s} p(s) \tag{1}$$

$$\pi_{kl} = \sum_{k \in s \wedge l \in s} p(s) \tag{2}$$

These inclusion probabilities are fundamental in the design-based school of inference. The inclusion probabilities are used in most estimators under the design-based framework.

In most cases you are interested in a property or variable of the objects observed. Let us denote this variable y. The entity of interest is mostly functions of the population as a whole or a subset, a "domain". These entities are called parameters and are denoted θ.

An example: If y is an indicator variable that indicates whether a respondent in the SCS has experienced robbery during the last year or not, a parameter of interest, denoted θ, would be the total number of people who have experienced robbery θ_1 or the proportion θ_2 , calculated in the following way (Särndal, Swensson and Wretman 1992, pp. 38-39):

$$\theta_1 = \sum_U y_i \tag{3}$$

$$\theta_2 = \frac{1}{N} \sum_U y_i \tag{4}$$

An estimator is a function of the sample, s that should be as close to the population parameter of interest as possible. The proportion of the sample that has experienced a certain offence can, for example, be used to estimate the proportion of the whole population. Since the sample is a random variable S, an estimator is a random variable of a given parameter where functions of random variables, as expectation and variance, can be calculated:

$$\hat{\theta} = \hat{\theta}(S) \tag{5}$$

$$E(\hat{\theta}) = \sum_{s \in \zeta} p(s)\hat{\theta}(s) \tag{6}$$

$$V(\hat{\theta}) = \sum_{s \in \zeta} p(s)\left\{\hat{\theta}(s) - E(\hat{\theta})\right\}^2 \tag{7}$$

To evaluate different design-based estimators both bias and variance are taken into account, and together they make up the mean squared error of the estimator. An estimator is considered design-unbiased if:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta = 0 \tag{8}$$

One of the most basic estimators for totals is the Horvitz-Thompson estimator for population totals. This estimator and the variance estimate are calculated the following way under simple random sampling (SRS):

$$\hat{t}_{HT} = \frac{N}{n}\sum_{k=1}^{n} y_k \tag{9}$$

$$\hat{V}(\hat{t}_{HT}) = N^2\left(1 - \frac{n}{N}\right)\sum_{k=1}^{n}\frac{(y_k - \bar{y})^2}{n(n-1)} \tag{10}$$

This estimator is both direct in the sense that the estimator only uses information from the sample. It is design-unbiased and does not use any auxiliary information in the estimation procedure apart from *N* (Särndal, Swensson and Wretman 1992, pp. 42-44).

One important characteristic of official statistics is that the confidence interval that is computed relies on approximately asymptotic unbiased estimators because unbiased estimators have interval estimates that can be interpreted correctly (Särndal, Swensson and Wretman 1992, pp. 40-41 and Lehtonen 2009, p. 15.).

In the area of small area estimation the main problem is that the domains where the estimates are sought for are often too small (by definition) to yield estimates with sufficient precision. Two main components that are needed to succeed in increasing the precision of small area estimates are good auxiliary information and a good linking model to make use of the auxiliary information. If those two components are present the variance of the estimates can be reduced for the estimates of interest (Rao 2003).

In this study the purpose is to find estimators that can be used in official statistics. In these estimates the estimates of interest is then often produced together with an interval estimate. This implies that interval estimates need to be based on design-unbiased estimators to be interpreted correctly, even though this assumption is complicated in the presence of survey non-response.

In the design-based framework, with its need of (asymptotically) design-unbiased estimators, the estimators used to incorporate auxiliary variables to reduce the estimator variance are so called model-assisted estimators. These estimators come in various shapes but are all, more or less, special cases of the generalized regression estimator (GREG). Examples of different GREG-estimators are the poststratification estimator, the ratio estimator and the regression estimator (Lehtonen and Pahkinen 2004, pp. 87-88.).

1.3.2.1. Design-based model-assisted small area estimators

The main type of estimators of interest in this study is the design-based model-assisted estimators and hence I will follow the work of Lehtonen and Pahkinen (2004).

These types of estimators come in various shapes, suited for different purposes. The first characteristic of the estimator you need to decide on is whether the estimator should be linear or nonlinear. Linear

estimators suit continuous response variables while nonlinear estimators, such as the logistic estimator, often work much better for binary data (Lehtonen and Pahkinen 2004, p. 196.).

The second property of the estimator is the covariate effects on the dependent variable. There are mainly three different ways that the auxiliary data can be included in the underlying model. The first model choice is the population-fixed effect where a model is fit to the whole population and the same coefficients in the model are constant over the domains. The second model choice is the domain-fixed effect model where domain effects are included in the model as slopes and/or intercepts in the model. The last model type is the mixed effect domain model, where there are both fixed effects and domain random effects (Lehtonen and Pahkinen 2004, pp. 196-198.).

The different types of estimators can be seen in table 2 (based on Lehtonen and Pahkinen 2004, p. 197.):

Table 2: Different model-assisted design-based small area estimators

Model effects	*Level of aggregation*	*Functional form*	*Estimator*	*Model*
Fixed effects	Population model	Linear	GREG-P	$E(y_k)= \boldsymbol{z}_k' \boldsymbol{\beta} + \varepsilon_k$
		Logistic	LGREG-P	$E(logit(y_k))= \boldsymbol{z}_k' \boldsymbol{\beta} + \varepsilon_k$
	Domain model	Linear	GREG-D	$E(y_k) = \boldsymbol{z}_k' \boldsymbol{\beta}_d + \varepsilon_k$
		Logistic	LGREG-D	$E(logit(y_k))= \boldsymbol{z}_k' \boldsymbol{\beta}_d + \varepsilon_k$
Mixed effects	Domain model	Linear	MGREG-D	$E(y_k)= \boldsymbol{z}_k'(\boldsymbol{\beta} + \boldsymbol{u}_d) + \varepsilon_k$
		Logistic	MLGREG-D	$E(logit(y_k))= \boldsymbol{z}_k'(\boldsymbol{\beta} + \boldsymbol{u}_d) + \varepsilon_k$

The main difference between the GREG-D and the GREG-P models is that the GREG-D model is direct in the sense that one model is produced for each domain, while the GREG-P model is indirect since the same coefficients in the model is used to estimate in all domains (se for example Lehtonen and Veijanen 2009 pp. 233-234).

One problem, especially with the domain specific models, is the problem of rare events. In many domains, especially when the domains get smaller and smaller, there will be situations with no events at all (all y will be zero, for instance). This means that there is a risk that the domain models will be very volatile and will not work very well.

When it comes to the specific problem of this study all the dependent variables will be binary. This means that the linear estimators are not suitable for this problem. Today in the SCS, the GREG-P estimator is used (as a calibration estimator). This estimator will be a reference point for the other estimators studied. The model is the model used to produce predicted values $\hat{y}_k$, in each domain, $\hat{y}_{kd} = z_k'\hat{\beta}$ in the linear case and $\hat{y}_{kd} = \frac{1}{1+e^{-z_k'\hat{\beta}}}$ in the logistic case.

All GREG-estimators of domain totals follow the same general estimation formula given by Lehtonen and Pahkinen (2004), pp. 198-200:

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} \frac{y_k - \hat{y}_k}{\pi_k} \tag{11}$$

where U_d is the population in domain d and s_d is the sample in the domain d.

An approximation of the variance of the estimators, given that the domains are planned, the direct estimator is used and the sampling design is simple random sampling in each domain, is given by the formula 6.15 in Lehtonen and Pahkinen 2004, p. 202.

$$\hat{V}(\hat{t}_{dGREG}) = N_d^2\left(1-\frac{n_d}{N_d}\right)\sum_{k\in s_d}\frac{(\hat{e}_k-\bar{\hat{e}}_d)^2}{n_d(n_d-1)} \tag{12}$$

where $\hat{e}_k = y_k - \hat{y}_k$ and $\bar{\hat{e}} = \sum_{k\in s_d}\frac{\hat{e}_k}{n_d}$.

There is a discussion whether the variance estimation should be calculated conditioned on the achieved sample size or over all theoretical samples (se for example Lehtonen and Pahkinen 2004, p. 202 and Holt and Smith 1979). In this study I will condition on the given sample, and hence all domain sample sizes will effectively be treated as planned. By conditioning on the sample size the variance is calculated over all theoretical samples with the same sample size, not over all samples.

To calculate an interval estimate for the parameter of interest the Wald confidence interval is then used (Särndal, Swensson and Wretman 1992, p. 238.).

$$\hat{t}_{dGREG} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{t}_{dGREG})} \tag{13}$$

The HT interval will be calculated the following way:

$$\hat{t}_{dHT} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{t}_{dHT})} \tag{14}$$

where $z_{1-\alpha/2}$ is the 1-α/2 percentile of the standard normal distribution.

Since the domains in these studies often are very small and the variance will be estimated the students t-distribution will be used instead of the standard normal distribution in interval estimation. The sample size (n) in each domain will be used as degrees of freedom.

1.3.3. Previous research – small area estimation in victimization surveys

A lot of research has been done in the area of small area estimation, both in the model-based and the Bayesian framework. An overview can be found in Rao (2003). Some research has also been made under the design-based framework, but the main focus has been on model-based and Bayesian approaches.

Small area estimation in victimization surveys has, until this date, only been studied in one small study in the Netherlands. In this study the authors, Buelens and Benschop, use a model-based approach to reduce the confidence intervals of violent crime victimization rates estimates. Estimates are produced for 25 different police zones in the Netherlands. These police zones were quite large compared to traditional small domains - the minimal number of respondents in each police zone was as large as 750 (Buelens and Benchop 2009, p.1.).

The model used by the authors is the Fay-Harriot (domain) area level model where victimization rates are modelled at the area level and with this model different auxiliary variables were evaluated as to how much they would reduce the variances of the estimates. By using both demographic auxiliary variables (proportion of people older than 30 years) and the number of reported violent crimes, the authors succeed in reducing the confidence interval with as much as 40 percent in some smaller domains. This shows the possible gains of using small area estimation in victimization surveys (Buelens and Benchop 2009, pp. 2-5.).

Even though no studies of design-based model-assisted estimators for small area estimation have been made in the area of victimization studies, some studies regarding model choice and model robustness of the GREG-family type estimators in small area estimation are of interest for this study.

In a Monte Carlo experiment by Lehtonen, Särndal and Veijanen (2005) different model structures were studied to see how these differences affect the properties of the estimators. One of the main conclusions of the study was that GREG estimators were only negligibly affected when it comes to accuracy of the estimator when mixed effect models were used instead of fixed effects models. The second conclusion was that neither domain specific slopes nor intercepts affected the accuracy of the estimators. Models with both domain specific slopes and intercepts were as good as the model that only included domain intercepts and only slightly better than the model without any domain specific information (Lehtonen, Särndal and Veijanen 2005, pp. 668–669.).

These results are positive since, with very rare events, there will be many domains with no occurrences at all. This will probably affect the robustness of the parameter estimates negatively more since a domain specific model is less robust than a population specific model.

Myrskylä (2007) studied the properties of the logistic regression estimator and came to the conclusion that the logistic GREG outperforms the linear GREG when the model is strong and the domain size is not very small ($n \geq 24$).

A general problem when it comes to rare events is estimation of confidence intervals of proportions from the binomial distribution. It has been shown by Brown, Cai and DasGupta (2001) that estimating a confidence interval of the binomial proportion with the classic normal-based Wald interval will result in confidence intervals with too low coverage probability, both regarding small samples and samples as large as n = 500. One of the main findings by Brown, Cai and DasGupta is that the Wilson interval outperforms the classical Wald intervals when it comes to very small proportions. This means that the Wilson interval will be studied and compared with the Wald interval based on the variance of the GREG estimates.

The classical Wilson confidence interval is calculated the following way:

$$\frac{\hat{p} + \frac{1}{2n} z_{1-\alpha/2}^2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n} z_{1-\alpha/2}^2} \tag{15}$$

where $\hat{p} = \frac{1}{n} \sum_{k \in s} y_k$ and $z_{1-\alpha/2}$ is the 1- α/2 percentile of the standard normal distribution.

As is obvious this interval is not motivated from a design-based perspective. Some design-based confidence intervals for small proportions have been suggested, but none of them for GREG estimators (see for example Korn and Graubard 1998).

A very simple way to use the Wilson interval for GREG estimates is simply to exchange the variance and the point estimates with the GREG point and variance estimates in the following way:

$$\frac{\hat{p}_{dGREG} + \frac{1}{2n_d} z_{1-\alpha/2}^2 \pm z_{1-\alpha/2} \sqrt{\hat{Var}(\hat{p}_{dGREG}) + \frac{z_{1-\alpha/2}^2}{4{n_d}^2}}}{1 + \frac{1}{n_d} z_{1-\alpha/2}^2} \tag{16}$$

where n_d is the total sample size for domain d for all strata. The Wilson interval will be calculated for the HT-estimator the following way:

$$\frac{\hat{p}_{dHT} + \frac{1}{2n_d} z^2_{1-\alpha/2} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_{dHT}(1-\hat{p}_{dHT})}{n_d} + \frac{z^2_{1-\alpha/2}}{4{n_d}^2}}}{1 + \frac{1}{n_d} z^2_{1-\alpha/2}} \tag{17}$$

Since the domains in these studies often are very small and the variance will be estimated the Students t-distribution will be used instead of the standard normal distribution in interval estimation. The sample size in each domain will be used as degrees of freedom.

1.4 Research questions

In this study two main questions will be discussed:

1) Which design-based estimators, point and interval, have the best properties for estimating victimization rates at county and municipality level in the Swedish crime survey?
2) How do point and interval estimates behave in the presence of rare events/skewed variables of interest?

1.5. The Swedish crime survey

In most countries there is an interest to know how the juridical system is working and how the safety of the citizens can be monitored and sustained. The traditional official statistics regarding crime and justice, such as reported crimes, convicted persons etc. have the problem that a large number of crimes are not reported to the police. These dark figures make it hard to estimate the ”true” number of crimes or offences based on the reported crimes, especially since different crimes have different dark numbers. Grand theft auto has a relatively small dark figures since the insurance companies demand a police report to pay the insurance claim while bicycle thefts has quite large dark figures (Irlander and Westfelt 2010:1, p. 44.).

To be able to estimate the number of offences or the number of victimized individuals many counties have complemented the regular crime statistics with a victimization survey. In different countries these types of studies have been conducted for different time periods. The National Victimization survey in the US, for example, has been conducted since 1973. In Sweden the Swedish victimization survey, The Swedish Crime Survey, was conducted in 2006 for the first time (see Groves et al 2004, p. 11 and Irlander and Westfelt 2010:1, p. 15.).

In the survey, a random sample of the Swedish population aged 16 - 79 years, is asked questions about experiences of offences (victimization), general safety, confidence in the Swedish justice system and victimized peoples’ contacts with the Swedish justice system. The mode of the survey is primarily computer-assisted telephone interviewing (CATI) but if the respondent does not want to answer by phone or is not reached, a paper questionnaire is sent to the respondent. The answers are complemented with register-based information such as age, sex, income etc (Irlander and Westfelt (2010:1), pp. 17-19.).

1.5.1. Sampling design, estimation and potential errors

The SCS is based on a stratified random sample of 20 000 people (with exception of the 2006 pilot survey where the sample size was 8 000) and has been conducted yearly since 2006. The frame is the Swedish registry for the total population created by Statistics Sweden. The stratification is made by demographic and geographic variables. Most sample sizes in strata are proportional to the population size but young males and small counties are over-sampled. The over-sampling counters low response rates among young people since these groups tend to respond less frequently than other groups, and the over sampling of small counties enables estimation at county level with sufficient precision (Irlander and Westfelt 2010:1, pp. 18-19.).

The percentage of survey non-respondents has increased during the years. The non-respondent rate has been around 20-30 percent during the years 2006-2009 and in the survey conducted in 2009 the number of non-respondent rose to 29.6 %, the highest level so far. To correct for non-response bias GREG-estimation (as calibration weight) is used for estimation. The same assisting model is then used both as national estimates and domain estimates.

1.6. Method

To evaluate the different estimators a design-based simulation study has been conducted. In this type of study a large number of samples is drawn from a defined population. Each estimate for each domain is then calculated and the properties of the estimators are evaluated over all these samples. This method, sometimes called Monte Carlo simulation or experiment, is one of the main methods used to evaluate different design-based estimators (see for example Särndal, Swensson and Wretman 1992, p.36 and Lehtonen and Pahkinen 2004, p. 210.).

1.6.1. Population, sample size, number of domains and sample design

The population that was used in this study is based on real life data. The respondents of the SCS during the years 2006 - 2009 were used with a total sample size of 50 514.[1] Partial non-response at the dependent variables of interest has been excluded from the study to facilitate the study. The size of the partial non-response was only 0,68 % of the population and for this reason the effect of excluding the partial non-response should not have any effects on the results of the study. The total size of the simulation population after exclusion of partial non-response was 50 173 respondents.

To minimize the effect of the finite population correction factor a sample cannot be the same size as the ordinary sample of 20 000. Other design-based simulation studies have used the proportion between sample and population of roughly 1 to 7 (Lehtonen and Pahkinen 2004, p. 210.). In this study the sample size is 5000, which is in the same proportion as other similar simulation studies but still the sample size is just one third of the overall received sample in one year from the SCS.

To be able to study the domain estimators, counties and municipalities will be clustered together 3 to 1 to resemble the sample collected sample size of the SCS. The clustering of municipalities and counties can be found in appendix 1.

The sample will be drawn with the same sampling design as in the SCS. A sample will be stratified and there will be at least 500 respondents in each cluster of 4 counties to resemble the real survey sampling design. In the study each domain will be estimated K times and the estimates for each domain will be studied depending on the mean sample size of the domain over all samples.

Since the SCS is stratified in two dimensions, by county and by age class, the age class strata will cut through all domains, even the very small domains. This means that situations will occur where the domains are so small that there will be only one or zero respondents in each stratum. In these cases the variance cannot be estimated and these domains will, for this reason, be excluded and not estimated.

1.6.2. Variables of interest to estimate

This study will evaluate the estimator properties for different response variables. Response variables that indicate experiences of assault, sexual offences, threat, robbery, fraud and harassment will be evaluated.

The reason for choosing different response variables is to study how different estimators perform under various levels of rare events. These different offences also differ with regard to dark figures. In 2008 only 19 percent of the sexual offence were reported to the police while as much as 43 percent of

1 The number of respondents and sample sizes has varied during the years. The number of respondents has been 13909 (2009), 14973 (2008), 14945 (2007) and 7687 (2006). For details see Irlander and Westfelt (2010:2), p. 18, Irlander and Wigerholt (2009), p.19, Töyrä and Wigerholt (2008), p. 17 and Töyrä (2007), p. 19.

the experienced robberies were reported (Irlander and Westfelt 2010:2, p. 28.). These dark figures should, theoretically, influence the effectiveness of using reported crimes as an auxiliary variable.

Table 3. Descriptive statistics of the response variables

Victimization rates			**Counties *d*=7**		**Municipalities *d*=97**	
Type	**N**	**Mean**	**Minimum**	**Maximum**	**Minimum**	**Maximum**
Robbery	51173	1,01%	0,49%	1,29%	0%	2,48%
Assault	51173	3,10%	2,48%	3,26%	0%	4,98%
Sexual offences	51173	0,93%	0,73%	1,00%	0%	2,13%
Threats	51173	4,58%	3,83%	4,88%	1,23%	9,76%
Fraud	51173	2,51%	1,83%	2,89%	0%	5,19%
Harassment	51173	4,51%	3,80%	4,99%	1,23%	9,25%
Any crime	51173	12,68%	10,35%	13,52%	6,13%	17,82%

You can see from the table that there is a difference between different municipalities and counties when it comes to how large proportion of the population have experienced the different crime types. In some municipalities the true proportion is that none has experienced the less frequently occurring offences such as robbery and sexual offences.

1.6.3. Choice of auxiliary variables

The results of Buelens and Benchop (2009) showed that the best auxiliary information in Holland for estimation of number of victims of violent crime was the number of reported violent crimes (Buelens and Benchop 2009, pp. 5-6). Other criminal offences, however, such as traffic offences and property crimes were not as good auxiliary variables in the model. These results will be used in the design of this study and in each model the number of reported crimes of the same type as the response variable will be included.

This approach will result in four sets of auxiliary variables used (see table 11 below), the auxiliary information used in the current SCS, the number of reported crimes of the same type together with the auxiliary variables used today and the number of reported crime, the interaction of the variable "if reported" number of reported crimes will be used and the auxiliary information used today. The last setup that will be used is the variable "have reported the crime to the police". This variable, today asked in the SCS, will be used as proxy for whether the respondent reported the crime at the unit level.

The number of reported crimes in the study will be calculated as the means of the number of reported crimes of the years included in the population. This is done since the population is based on different years with different numbers of reported crimes.

The crimes reported to the police follow the Swedish Penal law while the SCS ask questions that are less detailed than the Swedish penal code. In the SCS each question has a definition that makes it possible to compare the experienced offence with the reported number of crimes (Irlander and Westfelt 2010:1, pp. 25-26.). These definitions have been used in this study.

Table 4. Reported offences in the population

Reported offences		**Counties *d*=7**		**Municipalities *d*=97**	
per 100 000 citizens	**Mean**	**Minimum**	**Maximum**	**Minimum**	**Maximum**
Robbery	59	18	103	1	225
Assault	733	664	929	217	1192
Sexual offences	83	69	125	26	135
Threats	482	424	628	221	743

Fraud	286	307	614	106	725
Harassment	378	339	491	157	544

Table 5. Gender distribution in the population

Sex	**Frequency**	**Percent**
Male	24851	48,56
Female	26322	51,44

Table 6. Civil status in the population

Civil Status	**Frequency**	**Percent**
Not married	32021	62,57
Married or registered partnership	19152	37,43

Table 7. Age in the population

Age	**Frequency**	**Percent**
16-29	17613	34,42
30-40	7474	14,61
41-50	7138	13,95
51-65	10569	20,65
66-74	4566	8,92
75-79	3813	7,45

Table 8. Income in the population (in thousands)

Income (thousands of SEK)	**Frequency**	**Percent**
0-149	20765	40,58
150-299	21013	41,06
300-	9395	18,36

Table 9. The origin of the respondent

Born	**Frequency**	**Percent**
Sweden	47132	92,1
Elsewhere	4041	7,9

Table 10. Proportions living in urban/rural areas in the population

Municipality type	**Frequency**	**Percent**
Rural	34731	67,87
Urban	16442	32,13

Table 11. The different setups of auxiliary variables used in this study

Setup	**Variables**	**Comments**
Setup 1: SCS today	x_1 = county of residence x_3 = age class x_4 = born/not born in the Nordic countries	These auxiliary variables are currently used in the SCS. Here it is used as a baseline to compare with different variable setups.

	x_5 = civil status x_6 = income x_7 = urban/rural area x_8 = sex	
Setup 2: Reported crimes (area)	x_1 = county of residence x_3 = age class x_4 = born/not born in the Nordic countries x_5 = civil status x_6 = income x_7 = urban/rural area x_8 = sex x_9 = number of reported crimes in the municipality (of the same type)	As has been shown by Buelens and Benchop (2009), the number of reported crimes can explain a great part of the variation in victimization in an area level model. The question is whether this is possible when a unit-level model is used.
Setup 3: Municipality intercept	x_2 = municipality of residence x_3 = age class x_4 = born/not born in the Nordic countries x_5 = civil status x_6 = income x_7 = urban/rural area x_8 = sex	As has been shown by Lehtonen et al. (2005) a domain-based model is often slightly better than a population model. This means that this model should be better than the model used in the SCS today.
Setup 4: Reported crimes (unit level)	x_1 = county of residence x_3 = age class x_4 = born/not born in the Nordic countries x_5 = civil status x_6 = income x_7 = urban/rural area x_8 = sex x_{10} = if the respondent has reported the experienced crime	Buelens and Benchop (2009) showed that reported crimes at an area level could explain a great deal. In a unit level model, unit level information should carry more information and hence be much better to use than area level information. Variable x_{10} is an indicator variable that indicates if the respondent has reported a crime of the type that is estimated or not. If the victimization rate for harassment is estimated, the x_{10}-variable indicates if the respondent says he or she has reported harassment to the police during the year.

1.6.4. Estimators and models

Nine different estimators will be evaluated for each crime type, linear and logistic GREG for each of the four variable setups seen in table 11 above and the Horvitz-Thompson estimator. The Horvitz-Thompson estimator and the linear GREG-estimator used today (setup 1) in the SCS will be used as reference in the comparisons of the different estimators (Lehtonen and Pahkinen 2004, p. 199.).

All these estimators that will be studied are more or less based on the GREG-P model as can be seen from the table 11 above. Research has shown that there are very small gains to be had by borrowing strengths from other domains under the design-based framework (see for example Estevao and Särndal 2004). In setup 3 each municipality will have its own intercept but all other covariate effects will be estimated for the population as a whole. The reason why no GREG-D is used is because of the data structure with lot of rare binary dependent variables. A lot of municipalities will not have any occurrences of and hence the model will have problems estimating covariate coefficients. In this case it is simply not possible to use a GREG-D model.

Since the response variables are binary variables, logistic models are preferred (Lehtonen and Pahkinen 2004, pp. 196-197.). Logistic model estimators will use the same four setups of auxiliary variables.

Mixed-effect models will not be included in this study since earlier research concluded that the effect of a mixed-effect model should not be much better for modest sample sizes compared to a domain specific fixed effects model (Lehtonen, Särndal and Veijanen 2005, pp. 668–669.).

The estimation procedure will follow the operational steps given by Lehtonen and Pahkinen (2004). A model will be fit on the observed data after the sample is drawn. This model is then used to predict the value of each element that is not observed in the sample (the population). These results are then inserted in formula (11) and (12) to calculate the estimate and the variance of the estimate.

Interval estimates will be produced using the classical confidence (Wald) interval for GREG estimates (15) and HT estimates (16) and the Wilson interval based on GREG point and variance estimates (17) and the Wilson interval for HT estimates (19).

1.6.5. Number of simulated samples

To study the properties of the different estimators 1000 samples (1,2,…,*K*), will be drawn from the population. For each sample all estimates will be calculated and, based on these samples, measurements of performance for each estimator will be computed. This is approximately the same number of samples or more than similar simulation studies (See for example Lehtonen and Pahkinen 2004, p. 211.).

Since the smallest domain in some cases will be excluded (because n<2 in a stratum) the number of simulations will be smaller for the smallest domains.

1.6.6. Evaluation of the estimators

To evaluate the different estimators and to compare each estimator regarding bias and accuracy MSE and bias will be calculated over the simulated samples. Since there are some domains where the true total is zero, the absolute relative bias (ARB) and Relative root mean square error (RRMSE) cannot be used as a measure (See for example Lehtonen and Pahkinen 2004, p. 210.). Instead the average bias (AB) is used. This measure is calculated in the following way:

$$\left| \frac{1}{K} \sum_{v=1}^{K} \hat{p}_d(s_v) - p_d \right| \tag{18}$$

where $\hat{p}_d(s_v)$ is the estimate of the proportion of domain d for the simulated sample v and p_d is the true proportion in the domain d. To measure the accuracy of the estimator as a whole, Root mean square error (RMSE) will be used. RMSE is calculated in the following way:

$$\sqrt{\frac{1}{K} \sum_{v=1}^{K} (\hat{p}_d(s_v) - p_d)^2} \tag{19}$$

To be able to use the statistics in the production of official statistics interval estimates of good quality must be possible to produce, besides ordinary point estimates. Since the material is based on small samples and the response variable is highly skewed, any assumptions regarding normality of the estimates are probably not valid in the calculation of interval estimates, especially since it has been

shown by Brown, Cai and DasGupta (2001) that the classical Wald statistic for estimation of proportions is error prone. To study the properties of interval estimates (that is based on the Wald statistic) a coverage percent will be calculated for all estimators to evaluate the interval estimates.

Since some of the response variables (for example sexual offences) are very rare in the population, this means that, in some domains, there will be no occurrences of the variable of interest. In these cases some estimators (for example the HT-estimator) will not be able to compute variances or model parameters correctly and hence it will not be possible to compute interval estimates. This occurrence will be seen as an estimator breakdown and the percentage of estimator breakdowns will be used as a measurement of robustness for rare event small area estimation.

The rare event problem will also result in the risk of having calculated confidence interval, based on normality assumption that will be logically incorrect. Theoretically, in the presence of very skewed response variables, estimates will have confidence interval that will include negative values or values larger than one. This will be seen as logically incorrect confidence intervals. A percentage of the number of logically incorrect intervals will also be a measurement of the interval estimation quality.

Another problem that can arise with the GREG estimator is that the estimate can be negative. This can be problematic and research has been conducted on how to correct those problems. (see for example Stukel, Hidiroglou and Särndal 1996). The proportion of negative GREG-estimates for each estimator will be studied to see if there is a difference between the logistic and the linear GREG estimator. When the GREG estimate is negative, $\hat{p}$ is set to 0 when interval estimates are calculated.

1.6.7. Summary of technical details of Monte Carlo experiment

Population
N = 51 173
Respondents from the Swedish crime survey 2006 to 2009.
Elements with partial non-response excluded.

Sample size
n = 5000

Number of domains
D_1 = 7 large domains (counties)
D_2 = 97 small domains (municipalities).
See appendix 1 for details.

Target parameters:
Totals and/or proportions in each domain

Number of simulated samples:
K = 1000

Response variables
Experience during the last year:
y_1 = any crime type
y_2 = assault
y_3 = sexual offences
y_4 = threat
y_5 = robbery
y_6 = fraud
y_7 = harassment

Auxiliary data
x_1 = county of residence
x_2 = municipality of residence
x_3 = age class
x_4 = born/not born in the Nordic countries
x_5 = civil status
x_6 = income
x_7 = urban/rural area
x_8 = sex
x_9 = number of reported crimes (the same offences as the response variables above)
x_{10} = if the respondent has reported the experienced crime

Estimators:
$\hat{t}_{HT}$ = Horvitz-Thomson estimator (baseline),
$\hat{t}_{GREG-P}$ = GREG-estimator based on the different setups of auxiliary variables,
$\hat{t}_{LGREG-P}$ = logistic GREG-estimator based on the different setups of auxiliary variables,

Measures of performance:
Average bias (AB)
Root mean squared error (RMSE)
Estimator breakdown percent
Interval coverage percent
Logically incorrect interval percent

2. Results

2.1. Model fit of the GREG-estimators

To be able to produce small area estimates the degree of explanation made by the model is the key for good estimates. The better the model is to explain the variability in the data the better the estimators will perform. To study how well the different models work for the data all models have been fit to the whole population in the study to see how much different models explain of the variability.

Table 12: Model fit (R^2 linear regression), whole population

R^2-values (linear)	Setup 1	Setup 2	Setup 3	Setup 4
Robbery	0,8%	0,8%	0,9%	35,5%
Assault	3,1%	3,1%	3,2%	31,0%
Sexual offences	1,2%	1,2%	1,4%	16,2%
Threat	1,5%	1,5%	1,7%	20,0%
Fraud	0,6%	0,6%	0,8%	29,0%
Harassment	1,2%	1,2%	1,4%	14,7%
Any crime against persons	3,9%	4,0%	4,1%	26,1%

As you can see from the table above only a very small part of the variability is explained by the independent variables in the setup 1 to setup 3. This result means that since the model explains such a small part of the variance, the GREG should not perform much better than the ordinary HT-estimator. The fourth variable setup, on the other hand, where information on whether the person has reported the crime or not is included, seems to be a much better variable setup to use. From the table above it is obvious that when it comes to robbery and assault a quite large part of the variability in the data is explained by the variables in setup 4.

The differences between the crime types are probably explained by the different dark figures. Sexual offences and harassment, for example, have a larger dark figure than for example robbery (see Brå 2008, p.29) which means that a larger part of the variability in the data regarding robbery is explained by whether the person reported the crime to the police or not.

2.2. Point and variance estimator problems

One important aim of this study is to see how well the different estimators can handle rare events in the population (and in the sample). When it comes to rare events in the population there are some special situations in which it is not possible to produce credible estimates. This occurs in three situations as was explained in Chapter 1.6.6 above.

2.2.1. Proportions of possible variance estimates in the SCS

In the SCS today only counties are accounted for in the sampling design. This means that the municipality domain sample sizes can vary in size. Each domain, county and municipality, respectively, is also stratified by age in order to over sample young and elderly persons. In order to estimate the variance unbiasedly in the design-based framework each stratum needs a sample size of $n > 1$. This is not always the case for the smaller domains.

Table 13. Percentage of domains where all strata have a sample size $n > 1$

Mean sample size in domain	Number of domains	Percent (%)
-20	26	36.1
21 - 40	28	68.3
41 - 100	33	93.9

101 - 349	9	100.0
350 -	8	100.0

In the smallest domains, where the mean sample size is smaller than 20, the domains where the domain contain enough respondents in all strata's were only 36 per cent. The only way to cope with this problem is to include all domains in the sampling plan to be sure that all strata in all domains will have a sample size greater than 1.

This also means that the results in this study are based on a different number of simulated samples in different domains. But since the number of domains in the smallest mean sample size class is larger than in the largest class, the results should be equivalent and comparisons should be possible to do.

2.2.2. Variance estimates of zero for HT estimators

The first example of estimator problems for rare events is when the variance estimate is zero, which can happen for the HT variance estimator if there are only zeroes in the sample for a specific domain. In these cases it is not possible to calculate a logical Wald-interval estimate for the HT-estimator.

Table 14. Percentage of Horvitz-Thompson variance estimate of zero

Type of experienced offence	Mean sample size in domain				
	-20	**21 - 40**	**41 - 100**	**101 - 349**	**350 -**
Any crime against person	19.7	4.6	0.1	0.0	0.0
Assault	65.8	42.1	15.5	1.5	0.0
Fraud	73.3	54.0	24.2	4.0	0.0
Harassment	49.7	28.6	6.5	0.2	0.0
Robbery	93.8	79.7	55.4	22.9	4.2
Sexual offences	88.3	79.1	60.2	20.6	1.8
Threat	52.4	30.2	7.6	0.4	0.0

The table above shows that the proportions with a variance of zero get larger as the sample size gets smaller. The offences that are more rare, such as robbery and sexual offences also have a larger proportion of variance estimates of zero. This is logical and in line with what can be expected from theory.

2.2.3. Convergence of the logarithmic regression (LGREG)

Another problem is if the logarithmic regression does not converge. When it comes to the convergence the only problem was that in modelling Harassment, in 10 % of the simulations the ridging failed to improve the likelihood function when domain specific intercept were used. This should not affect the estimates to any larger extent and hence no way to improve the ridging was used.

2.2.4. Negative GREG estimates

The last situation where there can be problems with the point estimates is when the GREG point estimates are negative. This can happen with GREG estimators even though the results are logically wrong in the sense that negative proportions or totals cannot exist.

Table 15: Percentage of Negative GREG estimates

	Any crime	**Assault**	**Fraud**	**Harass-ment**	**Robbery**	**Sexual offences**	**Threat**

			against person						
Mean sample size in domain	**Auxiliary variables**	**Estimator type**							
-20	**Setup 1**	**Linear**	9.0	33.0	38.2	26.4	54.1	45.4	25.8
		Logistic	9.2	32.8	39.5	26.7	54.8	45.0	26.5
	Setup 2	**Linear**	8.7	33.1	37.4	26.5	53.6	44.5	25.4
		Logistic	9.1	32.8	39.6	26.7	55.0	44.9	26.5
	Setup 3	**Linear**	6.3	26.1	25.5	19.0	45.9	38.4	18.1
		Logistic	8.5	32.6	39.5	26.5	54.9	44.9	26.4
	Setup 4	**Linear**	5.1	21.5	27.7	17.9	51.2	39.9	15.2
		Logistic	5.0	21.5	27.9	17.9	52.1	39.7	15.2
21 - 40	**Setup 1**	**Linear**	1.5	17.3	24.5	11.9	39.4	36.6	12.2
		Logistic	1.5	17.0	24.8	11.8	39.1	35.3	12.2
	Setup 2	**Linear**	1.4	17.2	24.3	11.9	39.5	36.3	12.1
		Logistic	1.4	17.0	24.8	11.8	39.2	35.3	12.2
	Setup 3	**Linear**	1.3	16.5	22.4	11.2	37.6	34.8	11.4
		Logistic	1.2	17.0	24.8	11.9	39.2	35.5	12.4
	Setup 4	**Linear**	0.2	6.0	9.1	5.0	23.7	23.1	5.1
		Logistic	0.2	5.8	8.9	4.8	24.0	22.9	5.0
41 - 100	**Setup 1**	**Linear**	0.1	5.9	10.5	2.5	26.2	27.8	2.7
		Logistic	0.0	5.6	10.5	2.4	25.9	26.8	2.7
	Setup 2	**Linear**	0.1	5.9	10.4	2.5	26.2	27.7	2.7
		Logistic	0.0	5.6	10.4	2.4	26.0	26.8	2.7
	Setup 3	**Linear**	0.0	6.2	11.2	2.8	27.0	28.1	3.1
		Logistic	0.0	5.6	10.7	2.4	26.1	27.2	2.7
	Setup 4	**Linear**	0.0	0.9	1.3	0.9	6.2	15.6	0.6
		Logistic	0.0	0.9	1.2	0.9	6.0	15.3	0.5
101 - 349	**Setup 1**	**Linear**	0.0	0.6	1.6	0.0	10.2	8.8	0.1
		Logistic	0.0	0.5	1.7	0.0	9.9	8.1	0.1
	Setup 2	**Linear**	0.0	0.6	1.5	0.0	10.1	8.7	0.1
		Logistic	0.0	0.5	1.7	0.0	9.9	8.0	0.1
	Setup 3	**Linear**	0.0	0.6	1.9	0.1	10.6	9.3	0.1
		Logistic	0.0	0.5	1.7	0.0	10.2	8.2	0.1
	Setup 4	**Linear**	0.0	0.0	0.0	0.0	0.0	2.1	0.0
		Logistic	0.0	0.0	0.0	0.0	0.0	1.9	0.0
350 -	**Setup 1**	**Linear**	0.0	0.0	0.0	0.0	2.0	0.8	0.0
		Logistic	0.0	0.0	0.0	0.0	1.9	0.7	0.0
	Setup 2	**Linear**	0.0	0.0	0.0	0.0	2.0	0.9	0.0
		Logistic	0.0	0.0	0.0	0.0	1.9	0.8	0.0
	Setup 3	**Linear**	0.0	0.0	0.0	0.0	2.0	0.8	0.0
		Logistic	0.0	0.0	0.0	0.0	1.3	0.6	0.0
	Setup 4	**Linear**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Logistic	0.0	0.0	0.0	0.0	0.0	0.0	0.0

From the table above the conclusions can be drawn that setup 4 outperforms the other models when it comes to the proportion of negative estimates. In all domain sizes the proportion of negative estimates is highly reduced when information on whether the respondent reported the crime or not is used.

When it comes to differences between the logistic GREG and the linear GREG it is obvious that the logistic GREG is slightly better when the model is stronger (setup 4) and when the sample size is larger than 40. But the difference is quite small, even though the logistic estimator will only predict positive values.

Another interesting result is that in the more rare situations, such as robbery and sexual offences, there is still a quite large proportion of the estimates that are negative even for the larger domains. In the domains with a mean sample size of 101-349 still 10 percent of the estimates will be negative and hence erroneous.

2.3. Effectiveness of point estimates

The effectiveness of the estimator has been studied regarding both biases of the estimators as well as the mean squared error. The first property that is studied is the bias of the estimators:

Table 16: Average Bias (AB) of the estimated proportions

			Any crime against person	Assault	Fraud	Harass-ment	Robbery	Sexual offences	Threat
Mean sample size in domain	**Auxiliary variables**	**Estimator type**							
-20	**None**	**HT**	0.009	0.003	0.002	0.005	0.001	0.001	0.004
	Setup 1	**Linear**	0.005	0.003	0.002	0.004	0.001	0.001	0.003
		Logistic	0.005	0.003	0.002	0.004	0.001	0.001	0.003
	Setup 2	**Linear**	0.006	0.003	0.002	0.004	0.001	0.001	0.003
		Logistic	0.005	0.003	0.002	0.004	0.001	0.001	0.003
	Setup 3	**Linear**	0.006	0.002	0.002	0.004	0.001	0.001	0.003
		Logistic	0.005	0.002	0.002	0.004	0.001	0.001	0.003
	Setup 4	**Linear**	0.005	0.002	0.002	0.003	0.001	0.001	0.003
		Logistic	0.005	0.002	0.002	0.003	0.001	0.001	0.003
21 - 40	**None**	**HT**	0.002	0.001	0.001	0.001	0.000	0.001	0.001
	Setup 1	**Linear**	0.002	0.001	0.001	0.001	0.001	0.001	0.001
		Logistic	0.002	0.001	0.001	0.001	0.001	0.001	0.001
	Setup 2	**Linear**	0.002	0.001	0.001	0.001	0.001	0.001	0.001
		Logistic	0.002	0.001	0.001	0.001	0.001	0.001	0.001
	Setup 3	**Linear**	0.002	0.001	0.001	0.001	0.000	0.001	0.001
		Logistic	0.002	0.001	0.001	0.001	0.000	0.001	0.001
	Setup 4	**Linear**	0.002	0.001	0.001	0.001	0.000	0.000	0.001
		Logistic	0.002	0.001	0.001	0.001	0.000	0.000	0.001
41 - 100	**None**	**HT**	0.001	0.000	0.000	0.001	0.000	0.000	0.001
	Setup 1	**Linear**	0.001	0.000	0.000	0.001	0.000	0.000	0.001
		Logistic	0.001	0.000	0.000	0.001	0.000	0.000	0.001
	Setup 2	**Linear**	0.001	0.000	0.000	0.001	0.000	0.000	0.001
		Logistic	0.001	0.000	0.000	0.001	0.000	0.000	0.001
	Setup 3	**Linear**	0.001	0.000	0.000	0.001	0.000	0.000	0.001
		Logistic	0.001	0.000	0.000	0.001	0.000	0.000	0.001
	Setup 4	**Linear**	0.001	0.000	0.000	0.001	0.000	0.000	0.001
		Logistic	0.001	0.000	0.000	0.001	0.000	0.000	0.001
101 - 349	**None**	**HT**	0.001	0.001	0.000	0.001	0.000	0.000	0.000

	Setup 1	Linear	0.001	0.001	0.000	0.001	0.000	0.000	0.000
		Logistic	0.001	0.001	0.000	0.001	0.000	0.000	0.000
	Setup 2	Linear	0.001	0.001	0.000	0.001	0.000	0.000	0.000
		Logistic	0.001	0.001	0.000	0.001	0.000	0.000	0.000
	Setup 3	Linear	0.001	0.001	0.000	0.001	0.000	0.000	0.000
		Logistic	0.001	0.001	0.000	0.001	0.000	0.000	0.000
	Setup 4	Linear	0.001	0.000	0.000	0.001	0.000	0.000	0.000
		Logistic	0.001	0.000	0.000	0.001	0.000	0.000	0.000
350 -	None	HT	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Setup 1	Linear	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Logistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Setup 2	Linear	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Logistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Setup 3	Linear	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Logistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Setup 4	Linear	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Logistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000

As can be expected the bias is very small with these estimators - as theory suggest. From the results above you can see that all estimates can be regarded as design-unbiased.

Since all estimators are design-unbiased the RMSE is effectively the same as the variance of the estimators.

Table 17: Root mean squared error (RMSE) of the estimated proportions

			Any crime against person	Assault	Fraud	Harass-ment	Robbery	Sexual offences	Threat
Mean sample size in domain	**Auxiliary variables**	**Estimator type**							
-20	None	HT	0.083	0.039	0.034	0.054	0.011	0.017	0.051
	Setup 1	Linear	0.078	0.039	0.034	0.053	0.012	0.018	0.050
		Logistic	0.078	0.039	0.034	0.053	0.012	0.018	0.050
	Setup 2	Linear	0.078	0.039	0.034	0.053	0.012	0.018	0.050
		Logistic	0.078	0.039	0.034	0.053	0.012	0.018	0.050
	Setup 3	Linear	0.072	0.035	0.031	0.049	0.011	0.016	0.046
		Logistic	0.075	0.037	0.032	0.051	0.010	0.016	0.049
	Setup 4	Linear	0.070	0.033	0.030	0.049	0.011	0.016	0.044
		Logistic	0.070	0.033	0.030	0.049	0.011	0.016	0.045
21 - 40	None	HT	0.059	0.029	0.025	0.036	0.014	0.015	0.035
	Setup 1	Linear	0.054	0.028	0.025	0.035	0.014	0.015	0.034
		Logistic	0.054	0.028	0.025	0.035	0.014	0.015	0.034
	Setup 2	Linear	0.054	0.028	0.025	0.035	0.014	0.015	0.034
		Logistic	0.054	0.028	0.025	0.035	0.014	0.015	0.034
	Setup 3	Linear	0.054	0.028	0.025	0.035	0.014	0.015	0.034
		Logistic	0.055	0.029	0.025	0.036	0.014	0.016	0.034
	Setup 4	Linear	0.047	0.024	0.021	0.032	0.012	0.013	0.030
		Logistic	0.047	0.024	0.021	0.031	0.012	0.013	0.030
41 - 100	None	HT	0.042	0.021	0.018	0.025	0.011	0.010	0.025

	Setup 1	Linear	0.039	0.020	0.018	0.025	0.011	0.010	0.024
		Logistic	0.039	0.020	0.018	0.025	0.011	0.010	0.024
	Setup 2	Linear	0.039	0.020	0.018	0.025	0.011	0.010	0.024
		Logistic	0.039	0.020	0.018	0.025	0.011	0.010	0.024
	Setup 3	Linear	0.039	0.020	0.018	0.025	0.011	0.010	0.024
		Logistic	0.039	0.020	0.018	0.025	0.011	0.011	0.024
	Setup 4	Linear	0.034	0.017	0.015	0.023	0.008	0.009	0.021
		Logistic	0.034	0.017	0.015	0.023	0.008	0.009	0.021
101 - 349	None	HT	0.030	0.015	0.014	0.018	0.009	0.009	0.019
	Setup 1	Linear	0.028	0.015	0.014	0.017	0.009	0.009	0.018
		Logistic	0.028	0.015	0.014	0.017	0.009	0.009	0.018
	Setup 2	Linear	0.028	0.015	0.014	0.017	0.009	0.009	0.018
		Logistic	0.028	0.015	0.014	0.017	0.009	0.009	0.018
	Setup 3	Linear	0.028	0.015	0.014	0.017	0.009	0.009	0.018
		Logistic	0.028	0.015	0.014	0.017	0.009	0.009	0.018
	Setup 4	Linear	0.025	0.012	0.012	0.016	0.007	0.008	0.016
		Logistic	0.025	0.012	0.012	0.016	0.007	0.008	0.016
350 -	None	HT	0.014	0.007	0.006	0.009	0.004	0.004	0.009
	Setup 1	Linear	0.013	0.007	0.006	0.008	0.004	0.004	0.009
		Logistic	0.013	0.007	0.006	0.008	0.004	0.004	0.009
	Setup 2	Linear	0.013	0.007	0.006	0.008	0.004	0.004	0.009
		Logistic	0.013	0.007	0.006	0.008	0.004	0.004	0.009
	Setup 3	Linear	0.014	0.007	0.006	0.009	0.004	0.004	0.009
		Logistic	0.014	0.007	0.006	0.009	0.004	0.004	0.009
	Setup 4	Linear	0.012	0.006	0.005	0.008	0.003	0.004	0.008
		Logistic	0.012	0.006	0.005	0.008	0.003	0.004	0.008

From the table above you can see that using the variable setup 4 means quite a large gain compared to using the other setups, especially in the smaller domains and when the criminal offence is less common. To clarify the gains in estimator precision the following table compares the RMSE between Setup 1 and 4 for the linear GREG estimator:

Table 18: RMSE ratio between GREG (linear) Setup 4 and GREG (linear) Setup 1 from table X above

Mean sample size in domain	Any crime against person	Assault	Fraud	Harass-ment	Robbery	Sexual offences	Threat
-20	0.901	0.851	0.881	0.926	0.859	0.905	0.895
21 - 40	0.868	0.860	0.852	0.892	0.824	0.850	0.894
41 - 100	0.874	0.840	0.841	0.934	0.756	0.911	0.890
101 - 349	0.885	0.851	0.864	0.936	0.814	0.929	0.905
350 -	0.878	0.847	0.842	0.924	0.778	0.925	0.901
All	0.881	0.850	0.856	0.923	0.806	0.904	0.897

As expected, the gains in precision are larger for robbery and assault and smaller for sexual offences and harassment. The crime types as sexual offences and harassment that has larger dark figures have a worse model fit compared to assault, that is recognized as having a smaller dark figure. As such the estimator precision gains for setup 4 is, as expected, larger for the crime types with a smaller dark figure.

The gain of using setup 4 varies between the crime types but the RMSE is reduced both for harassment and sexual offences as well as for estimates for assault and robbery. To get the same gains by increasing the sample size one would have to increase the sample size by between 8 percent (harassment) and 24 percent (robbery).

2.4. Interval estimates

As has been concluded in the section purpose of this study, an important part of the production of official statistics is credible confidence intervals for different domains of interest. The first interval is the classical Wald interval that is usually applied for the GREG and HT estimators.

2.4.1. Wald (GREG) interval estimator

Table 19: Wald coverage proportion (95%, based on Student's t-distribution)

			Any crime against person	Assault	Fraud	Harass-ment	Robbery	Sexual offences	Threat
Mean sample size in domain	**Auxiliary variables**	**Estimator type**							
-20	**None**	**HT**	0.760	0.332	0.254	0.487	0.060	0.110	0.461
	Setup 1	**Linear**	0.769	0.444	0.369	0.515	0.602	0.510	0.506
		Logistic	0.769	0.455	0.350	0.512	0.597	0.543	0.502
	Setup 2	**Linear**	0.769	0.445	0.368	0.515	0.602	0.507	0.504
		Logistic	0.769	0.455	0.347	0.513	0.593	0.516	0.495
	Setup 3	**Linear**	0.784	0.442	0.373	0.512	0.603	0.513	0.498
		Logistic	0.785	0.406	0.294	0.498	0.574	0.431	0.471
	Setup 4	**Linear**	0.710	0.382	0.343	0.486	0.634	0.494	0.456
		Logistic	0.709	0.397	0.328	0.483	0.633	0.528	0.453
21 - 40	**None**	**HT**	0.874	0.569	0.456	0.704	0.201	0.207	0.689
	Setup 1	**Linear**	0.887	0.591	0.471	0.710	0.347	0.361	0.695
		Logistic	0.886	0.599	0.469	0.711	0.341	0.435	0.696
	Setup 2	**Linear**	0.887	0.592	0.470	0.710	0.346	0.358	0.695
		Logistic	0.886	0.598	0.468	0.711	0.335	0.393	0.695
	Setup 3	**Linear**	0.897	0.585	0.463	0.711	0.337	0.352	0.697
		Logistic	0.895	0.577	0.459	0.711	0.295	0.207	0.696
	Setup 4	**Linear**	0.855	0.487	0.424	0.651	0.356	0.362	0.614
		Logistic	0.854	0.499	0.418	0.652	0.363	0.435	0.615
41 - 100	**None**	**HT**	0.914	0.816	0.751	0.876	0.444	0.396	0.869
	Setup 1	**Linear**	0.928	0.820	0.751	0.880	0.506	0.447	0.872
		Logistic	0.929	0.821	0.752	0.880	0.488	0.472	0.872
	Setup 2	**Linear**	0.928	0.820	0.752	0.881	0.505	0.447	0.872
		Logistic	0.929	0.821	0.752	0.880	0.487	0.468	0.872
	Setup 3	**Linear**	0.931	0.821	0.753	0.883	0.503	0.440	0.876
		Logistic	0.931	0.822	0.753	0.883	0.471	0.396	0.876
	Setup 4	**Linear**	0.917	0.730	0.641	0.865	0.410	0.396	0.837
		Logistic	0.917	0.732	0.641	0.865	0.396	0.422	0.838
101 - 349	**None**	**HT**	0.925	0.901	0.903	0.918	0.750	0.777	0.918
	Setup 1	**Linear**	0.942	0.904	0.905	0.926	0.751	0.780	0.921
		Logistic	0.942	0.903	0.904	0.926	0.750	0.781	0.921

	Setup 2	Linear	0.942	0.904	0.904	0.925	0.751	0.781	0.921
		Logistic	0.942	0.903	0.904	0.926	0.750	0.781	0.921
	Setup 3	Linear	0.944	0.904	0.906	0.925	0.751	0.780	0.923
		Logistic	0.942	0.904	0.904	0.924	0.751	0.779	0.922
	Setup 4	Linear	0.938	0.892	0.885	0.920	0.633	0.737	0.912
		Logistic	0.938	0.894	0.885	0.922	0.632	0.737	0.911
350 -	None	HT	0.942	0.934	0.931	0.942	0.922	0.890	0.936
	Setup 1	Linear	0.945	0.937	0.927	0.940	0.920	0.890	0.938
		Logistic	0.944	0.935	0.927	0.940	0.920	0.887	0.938
	Setup 2	Linear	0.945	0.937	0.927	0.940	0.920	0.890	0.938
		Logistic	0.944	0.935	0.927	0.940	0.920	0.888	0.938
	Setup 3	Linear	0.941	0.934	0.923	0.937	0.916	0.889	0.934
		Logistic	0.939	0.930	0.921	0.937	0.912	0.885	0.934
	Setup 4	Linear	0.943	0.932	0.920	0.939	0.838	0.897	0.935
		Logistic	0.943	0.932	0.920	0.939	0.839	0.898	0.935

From the results in the table above one the conclusion can be drawn that the Wald interval does not work at all for domain estimation for more rare events. For proportions that are small, such as sexual offences and robbery, the coverage proportions are extremely low even for the large domains (101-349). But for larger proportions such as "Any crime against person" the Wald interval works quite well for even for medium sized domains such as 41 - 100.

The Wald interval is based on the assumption on the central limit theorem approximation of the normal distribution. To study this assumption the proportions of intervals where values smaller than 0 (or greater than 1) is covered by the interval is calculated. These logically incorrect intervals suggest that the asymptotic of the central limit theorem does not work empirically for this sample size and can explain the coverage problems seen above.

Table 20: Wald incorrect interval proportions (95%, based on Student's t-distribution)

			Any crime against person	Assault	Fraud	Harass-ment	Robbery	Sexual offences	Threat
Mean sample size in domain	**Auxiliary variables**	**Estimator type**							
-20	None	HT	0.554	0.306	0.237	0.423	0.057	0.106	0.402
	Setup 1	Linear	0.761	0.850	0.861	0.873	0.903	0.919	0.834
		Logistic	0.759	0.874	0.846	0.868	0.894	0.911	0.833
	Setup 2	Linear	0.756	0.850	0.864	0.874	0.904	0.917	0.835
		Logistic	0.755	0.874	0.847	0.869	0.893	0.912	0.832
	Setup 3	Linear	0.782	0.858	0.897	0.915	0.915	0.926	0.861
		Logistic	0.767	0.885	0.860	0.889	0.894	0.913	0.854
	Setup 4	Linear	0.605	0.600	0.689	0.739	0.850	0.812	0.645
		Logistic	0.600	0.621	0.674	0.733	0.843	0.819	0.643
21 - 40	None	HT	0.495	0.510	0.428	0.597	0.193	0.202	0.594
	Setup 1	Linear	0.526	0.834	0.859	0.836	0.867	0.907	0.825
		Logistic	0.525	0.856	0.842	0.832	0.869	0.929	0.826
	Setup 2	Linear	0.525	0.834	0.863	0.837	0.873	0.910	0.825
		Logistic	0.525	0.856	0.842	0.832	0.870	0.929	0.824
	Setup 3	Linear	0.517	0.868	0.931	0.893	0.908	0.927	0.875

		Logistic	0.511	0.871	0.852	0.840	0.874	0.927	0.836
	Setup 4	**Linear**	0.263	0.537	0.573	0.644	0.576	0.669	0.614
		Logistic	0.261	0.550	0.553	0.643	0.591	0.705	0.610
41 - 100	**None**	**HT**	0.125	0.625	0.635	0.569	0.425	0.386	0.580
	Setup 1	**Linear**	0.083	0.755	0.840	0.626	0.888	0.918	0.645
		Logistic	0.083	0.759	0.832	0.625	0.885	0.938	0.644
	Setup 2	**Linear**	0.083	0.755	0.840	0.627	0.891	0.919	0.645
		Logistic	0.083	0.758	0.831	0.625	0.885	0.938	0.644
	Setup 3	**Linear**	0.086	0.770	0.882	0.634	0.930	0.947	0.653
		Logistic	0.087	0.758	0.836	0.618	0.890	0.940	0.637
	Setup 4	**Linear**	0.010	0.408	0.485	0.489	0.425	0.701	0.405
		Logistic	0.009	0.409	0.478	0.485	0.420	0.729	0.404
101 - 349	**None**	**HT**	0.001	0.380	0.475	0.162	0.605	0.671	0.159
	Setup 1	**Linear**	0.000	0.366	0.503	0.132	0.805	0.860	0.130
		Logistic	0.000	0.361	0.499	0.133	0.797	0.858	0.130
	Setup 2	**Linear**	0.000	0.366	0.501	0.132	0.803	0.858	0.129
		Logistic	0.000	0.361	0.498	0.132	0.797	0.857	0.129
	Setup 3	**Linear**	0.000	0.365	0.504	0.137	0.820	0.871	0.134
		Logistic	0.000	0.356	0.499	0.138	0.797	0.851	0.135
	Setup 4	**Linear**	0.000	0.063	0.217	0.049	0.382	0.688	0.027
		Logistic	0.000	0.064	0.216	0.049	0.379	0.676	0.027
350 -	**None**	**HT**	0.000	0.002	0.028	0.000	0.434	0.344	0.000
	Setup 1	**Linear**	0.000	0.002	0.027	0.000	0.469	0.353	0.000
		Logistic	0.000	0.002	0.028	0.000	0.469	0.347	0.000
	Setup 2	**Linear**	0.000	0.002	0.027	0.000	0.469	0.353	0.000
		Logistic	0.000	0.002	0.028	0.000	0.468	0.348	0.000
	Setup 3	**Linear**	0.000	0.002	0.028	0.000	0.456	0.336	0.000
		Logistic	0.000	0.003	0.027	0.000	0.443	0.322	0.000
	Setup 4	**Linear**	0.000	0.000	0.000	0.000	0.087	0.200	0.000
		Logistic	0.000	0.000	0.000	0.000	0.088	0.202	0.000

As can be seen from the table above the proportion of incorrect intervals decreases both with the sample size and the rareness in the populations – which is in line with the central limit theorem. It can also be seen that the variable setup 4 outperforms the other setups with auxiliary variables.

When it comes to rare events as sexual offences and robbery the interval does not converge to the normal distribution even in the largest domains. Using the variables used today in the SCS only 55 % of the estimated confidence intervals would be correct in the sense that they wouldn't include negative values.

As concluded above the coverage problems and the inclusion of negative values in the confidence interval suggest that the distribution of the estimators does not converge to the normal distribution and hence the Wald interval should not be used to estimate confidence intervals for rare events in domains, even in very large domains.

There is no big difference between the logistic and the linear estimator in these cases. The logistic estimator is slightly better than the linear, but the difference is very small.

2.4.2. Wilson intervals

The Wilson interval is the interval recommended by Brown et al (2001) for estimating a binomial proportion. In this study the interval has been applied *ad hoc* to the design-based framework in Chapter 1.3.2 above and in the study the Students t-distribution has been used in the interval estimation to include the uncertainty of the variance estimate in the study.

Table 21: Wilson ad hoc interval coverage proportion (95%, based on Student's t-distribution)

			Any crime against person	Assault	Fraud	Harass-ment	Robbery	Sexual offences	Threat
Mean sample size in domain	**Auxiliary variables**	**Estimator type**							
-20	**None**	**HT**	0.947	0.926	0.920	0.942	0.742	0.820	0.951
	Setup 1	**Linear**	0.945	0.887	0.881	0.922	0.719	0.776	0.913
		Logistic	0.946	0.886	0.883	0.921	0.722	0.775	0.913
	Setup 2	**Linear**	0.944	0.887	0.877	0.922	0.717	0.770	0.912
		Logistic	0.945	0.886	0.880	0.922	0.722	0.774	0.912
	Setup 3	**Linear**	0.960	0.900	0.902	0.941	0.678	0.759	0.933
		Logistic	0.957	0.888	0.891	0.935	0.597	0.715	0.930
	Setup 4	**Linear**	0.944	0.883	0.890	0.918	0.724	0.777	0.914
		Logistic	0.944	0.880	0.891	0.919	0.720	0.774	0.914
21 - 40	**None**	**HT**	0.954	0.960	0.960	0.957	0.897	0.940	0.959
	Setup 1	**Linear**	0.956	0.930	0.920	0.942	0.854	0.887	0.945
		Logistic	0.955	0.928	0.920	0.941	0.854	0.881	0.945
	Setup 2	**Linear**	0.955	0.930	0.920	0.942	0.853	0.888	0.945
		Logistic	0.955	0.929	0.920	0.942	0.854	0.884	0.945
	Setup 3	**Linear**	0.964	0.937	0.929	0.951	0.858	0.899	0.952
		Logistic	0.961	0.930	0.924	0.945	0.829	0.899	0.948
	Setup 4	**Linear**	0.959	0.925	0.920	0.940	0.837	0.865	0.943
		Logistic	0.959	0.925	0.920	0.940	0.838	0.860	0.943
41 - 100	**None**	**HT**	0.953	0.966	0.962	0.963	0.940	0.956	0.965
	Setup 1	**Linear**	0.955	0.946	0.947	0.955	0.907	0.914	0.956
		Logistic	0.954	0.946	0.947	0.955	0.908	0.912	0.955
	Setup 2	**Linear**	0.954	0.946	0.947	0.955	0.907	0.914	0.956
		Logistic	0.954	0.946	0.946	0.955	0.908	0.912	0.955
	Setup 3	**Linear**	0.957	0.950	0.948	0.957	0.909	0.917	0.958
		Logistic	0.956	0.947	0.947	0.955	0.901	0.910	0.956
	Setup 4	**Linear**	0.958	0.944	0.944	0.955	0.882	0.909	0.956
		Logistic	0.958	0.943	0.944	0.955	0.888	0.905	0.956
101 - 349	**None**	**HT**	0.951	0.961	0.967	0.965	0.958	0.965	0.960
	Setup 1	**Linear**	0.952	0.947	0.957	0.957	0.942	0.948	0.952
		Logistic	0.952	0.947	0.957	0.956	0.942	0.942	0.952
	Setup 2	**Linear**	0.952	0.947	0.957	0.956	0.942	0.948	0.952
		Logistic	0.952	0.947	0.957	0.956	0.942	0.941	0.952
	Setup 3	**Linear**	0.954	0.948	0.959	0.956	0.941	0.945	0.951
		Logistic	0.955	0.947	0.956	0.952	0.940	0.935	0.951
	Setup 4	**Linear**	0.953	0.953	0.954	0.958	0.940	0.940	0.953
		Logistic	0.952	0.953	0.952	0.957	0.939	0.940	0.954
350 -	**None**	**HT**	0.959	0.964	0.960	0.961	0.967	0.957	0.961
	Setup 1	**Linear**	0.949	0.948	0.951	0.951	0.955	0.950	0.948
		Logistic	0.948	0.946	0.951	0.952	0.954	0.945	0.948
	Setup 2	**Linear**	0.949	0.948	0.951	0.951	0.955	0.949	0.947

		Logistic	0.949	0.947	0.951	0.952	0.954	0.945	0.948
	Setup 3	**Linear**	0.945	0.944	0.946	0.947	0.948	0.943	0.944
		Logistic	0.945	0.940	0.945	0.947	0.944	0.931	0.941
	Setup 4	**Linear**	0.950	0.945	0.952	0.952	0.952	0.946	0.947
		Logistic	0.949	0.944	0.952	0.951	0.950	0.948	0.947

From the table you can see that the Wilson interval performs much better than the Wald interval in almost all situations. One example is estimates for sexual offences and robbery in the 101-349 domain sizes where the Wilson interval has a coverage of between 94 and 96 percent while the Wald interval has a coverage between 63 and 77 percent. When it comes to the more common variable "Any crime against person", the Wilson interval again has a coverage percentage between 94 and 96 percent while the Wald interval has a coverage of only 71 and 78 percent. Clearly the Wilson interval outperforms the Wald interval in these situations.

In the larger domains (100 -) the Wilson ad hoc interval performs generally very well both for the HT estimator as well as the GREG estimators. But in the smaller domains the Wilson interval have a coverage proportion that can be considered good only for the crime types that are less rare.

The reason may be that the true values in many of those domains are zero and the Wilson interval only very seldom include zeros in the interval. If the population would be larger (as in the SCS) the probability that the municipality would have zero occurrences would be much smaller and hence the interval coverage would perform better.

Table 22: Wilson ad hoc interval coverage proportion (95%, based on Student's t-distribution) for more rare offences

**= no domains where the population value in the domain is zero*

			Robbery		**Sexual offences**	
			True value in population		**True value in population**	
			> 0	**0**	**> 0**	**0**
Mean sample size in domain	**Auxiliary variables**	**Estimator type**				
-20	**None**	**HT**	0.898	0.488	0.916	0.465
	Setup 1	**Linear**	0.902	0.582	0.893	0.546
		Logistic	0.901	0.583	0.889	0.546
	Setup 2	**Linear**	0.901	0.580	0.892	0.531
		Logistic	0.901	0.584	0.892	0.542
	Setup 3	**Linear**	0.920	0.498	0.915	0.457
		Logistic	0.917	0.300	0.908	0.261
	Setup 4	**Linear**	0.890	0.595	0.896	0.539
		Logistic	0.889	0.585	0.891	0.535
21 - 40	**None**	**HT**	0.940	0.525	0.934	*
	Setup 1	**Linear**	0.888	0.546	0.883	*
		Logistic	0.888	0.531	0.877	*
	Setup 2	**Linear**	0.887	0.545	0.885	*
		Logistic	0.887	0.534	0.881	*
	Setup 3	**Linear**	0.898	0.497	0.896	*
		Logistic	0.896	0.231	0.896	*
	Setup 4	**Linear**	0.868	0.549	0.860	*

		Logistic	0.873	0.503	0.856	*
41 - 100	None	HT	0.950	0.504	0.952	*
	Setup 1	Linear	0.917	0.522	0.912	*
		Logistic	0.918	0.526	0.909	*
	Setup 2	Linear	0.917	0.529	0.912	*
		Logistic	0.918	0.526	0.910	*
	Setup 3	Linear	0.919	0.515	0.915	*
		Logistic	0.918	0.281	0.908	*
	Setup 4	Linear	0.891	0.516	0.906	*
		Logistic	0.898	0.488	0.903	*

From the table you can see that the main problem for the interval estimates is when the true value is zero, which should not be a real problem in the SCS. But even if the true value is considered to be larger than zero the Wilson interval has problems with the coverage probability for these very rare crime types. But compared to the Wald interval the Wilson interval is to be preferred even in these situations.

If the HT-interval estimates are considered the Wilson interval is also to be perferred. The problem in small domains and rare events is that the variance estimate is zero and the Wald interval cannot be computed. By using the Wilson interval instead, interval estimates with good coverage proportions can be estimated.

Even though it would be possible to produce interval estimates for even the smallest domain they may not be results of any interests for the users. The interval width in the smallest domains may be so large that the interval estimates are probably of no real use.

2.4.3. Comparing the width of the interval estimates

Above the different interval estimators have been studied regarding the coverage proportions of the estimates. Another important aspect of the interval estimate is the width of the estimates. Too wide estimates will not be of any use. If the width of the interval was mere a function of the variance the RMSE in Chapter 2.3 above would be the end of the story. But the Wilson interval in not just the function of the variance, but also a function of the domain size.

Table 23: Interval mean width (95%, based on Student's t-distribution)

			Any crime against person		Assault		Harassment		Robbery		Sexual offences	
			Wilson	Wald	Wilson	Wald	Wilson	Wald	Wilson	Wald	Wilson	Wald
Mean sample size in domain	Auxiliary variables	Estimator type										
-20	None	HT	33.9	28.9	26.3	9.5	28.3	15.2	23.7	1.6	24.1	2.9
	Setup 1	Linear	34.0	29.5	26.3	10.5	28.3	16.1	23.7	2.2	24.1	3.6
	Setup 4	Linear	32.1	25.3	25.5	7.8	27.7	14.4	23.6	1.8	24.0	3.1
21 - 40	None	HT	23.7	22.0	16.5	9.0	18.0	12.1	13.9	2.7	13.9	2.9
	Setup 1	Linear	23.4	22.0	16.4	9.4	17.8	12.4	13.8	3.0	13.9	3.3
	Setup 4	Linear	21.3	18.6	15.5	7.3	17.0	10.7	13.6	2.3	13.6	2.7
41 - 100	None	HT	16.5	15.5	10.0	7.3	11.4	9.5	7.5	3.0	7.3	2.6
	Setup 1	Linear	15.8	15.5	9.7	7.4	11.1	9.5	7.4	3.1	7.3	2.9
	Setup 4	Linear	14.1	13.4	8.8	5.8	10.5	8.7	7.0	2.1	7.1	2.4

101 - 349	**None**	**HT**	12.0	11.3	6.6	5.7	7.7	6.9	4.7	3.1	4.6	3.1
	Setup 1	**Linear**	11.3	11.2	6.3	5.7	7.4	6.9	4.5	3.1	4.5	3.1
	Setup 4	**Linear**	10.0	9.9	5.6	4.7	7.0	6.4	4.1	2.3	4.3	2.9
350 -	**None**	**HT**	5.7	5.3	3.0	2.7	3.6	3.3	1.7	1.4	1.8	1.5
	Setup 1	**Linear**	5.3	5.3	2.8	2.7	3.4	3.3	1.7	1.4	1.7	1.5
	Setup 4	**Linear**	4.6	4.6	2.4	2.3	3.2	3.1	1.4	1.1	1.6	1.4

From the table above you can see that the Wilson interval enlarges the interval for the smallest domains. Compared to the Wald interval the difference can be extremly large in the smallest domains. Another interesting feature of these estimators is that the effect of using setup 4 to reduce the variance of the estimators is larger as the domain size gets larger. When it comes to sexual offences and harassment using setup 4 would mean only a very small gain.

3. Conclusions

In this study the following two research question were considered:

1) Which design-based estimator, point and interval, have the better properties for estimating victimization rates at county and municipality level in the Swedish crime survey?
2) How do point and interval estimates behave in the presence of rare events/skewed variables of interest?

Based on earlier research two types of point estimators were included in this study. The first one was the GREG estimator, which is currently used in the SCS. The second type was the logistic GREG that has been proven to work better in some situations than the ordinary GREG.

In the previous research regarding small area estimates for victimization surveys reported crimes have been identified as good auxiliary variables to model victimization. Criminological theory suggests that this information should work better for crime types that have a smaller dark figure. For this reason and since earlier research suggests that domain specific intercepts may improve estimates different variable setups were examined in this study. The auxiliary variables used today, reported crimes at the area level (municipality), domain specific intercepts and whether the respondent reported the crime or not to the police was studied

Regarding interval estimates the classical Wald interval was included in the study. As a complement to the Wald interval that, by previous research, had been shown to perform very poorly, the Wilson interval was adapted *ad hoc* to the design-based framework and the GREG estimator.

3.1. Auxiliary variables

As already concluded in Chapter 2.1 the different variable setups differ in the explained variance. Variable setup 4 explains a lot more of the variance than the other variable setups. As suspected the explained variances differ between the crime types where robbery has a larger part of explained variability than sexual offences.

3.2. GREG and LGREG type estimators

The results in this study show that the difference in bias between the LGREG and the GREG estimator is very small. Even if the LGREG tends to have a larger bias in the smallest domains, the difference is very small between the two estimators.

When it comes to the RMSE the LGREG estimator is slightly better than the GREG estimator. But, as with the bias, the difference is very small ($n \geq 24$). Earlier research by Myrskylä (2007) showed that the logistic GREG could outperform the GREG if the model was strong. In the case with variable setup 4, a R^2-value of 0.30 is apparently not strong enough to give sufficient gains in the estimation.

The problem with negative estimates is slightly smaller with the LGREG the GREG estimator, but the difference is to small to be of any practical use. The skewness of the dependent variables has a huge impact on both the GREG estimates and the LGREG estimates when it comes to negative estimates. The more skewed the variable is, and the smaller the sample, the more negative estimates. “Any crime against persons” with a true *p* of 12 % in the population only gets 7 % negative values in the smallest domains while when it comes to more rare events such as sexual offences more than 40 % of the estimates were negative in the smallest domains. The main difference depends on the strength of the auxiliary variables. With variable setup 4 which can be considered as a stronger model the proportion of negative estimates is reduced by as much as a half in some situations.

3.3. Interval estimation

The main difference between the Wald and Wilson interval is that the Wilson interval takes the sample size explicitly into account when estimating the confidence interval – to correct for small (or non-existing) variance estimates of the binomial proportion. The Wald interval, on the other hand, only takes the variance estimate of the estimator into account and that the estimator is assumed to be normally distributed – given by the central limit theorem.

This means that the Wald interval is always shorter than the Wilson interval and from that point of view the Wald interval is to be preferred. The main problem is that the Wald interval relies on the central limit theorem to approximate to the normal distribution, an assumption that does not always hold. In the situation with small samples and skewed variables the convergence is not fast enough.

Theory suggests that the skewer the variable the longer it takes to converge to the normal distribution. In this study the conclusion is that for the more rare events in the population ($p < 1\ \%$) the Wald interval cannot be used at all, even in the largest domains, while for the less skewed variable "Any crimes against persons", the Wald interval can be used for sample sizes down to $n>100$. This is of course dependent on how strict you are when it comes to the coverage proportion of the estimates.

The Wilson interval works quite well for almost all domains. For the larger domains ($n>100$) the coverage proportions are around 95%. But the skewer the variable and the smaller the domain, the worse the coverage probability will be. But for the more rare offences as robbery and sexual offences the coverage proportions are quite low, just around 90%, in the domains that are smaller than 100. But still the Wilson interval is much better than the Wald interval in these situations as well.

One main problem with the Wilson interval is that when the true value in the domain is zero, the Wilson interval gets very erratic convergence probability. This is not a huge problem for the estimating intervals in practical situations if the population is sufficiently large, but for smaller population it can be a problem. On the other hand, though, if you suspect that the population value may be zero you may consider a one-sided interval instead. This has not been examined in this study.

Since the difference between the Wald interval and the Wilson interval gets smaller as the domain size gets larger, the interval estimator that should be used in all situations should be the Wilson interval.

3.4. Suggestions for the SCS

The results from this study can be divided into four different conclusions regarding production and dissemination of statistics for the Swedish crime survey.

3.4.1. Using reported crime at the unit level should be investigated further

The study clearly shows that using the variable whether a person has reported a crime or not can improve the estimates. The two main benefits are the reduction of negative GREG estimates as well as the reduction in interval width for some variables. Together with the possibility to study the measurement errors created by telescoping effects this is an area that should be investigated further.

3.4.2. The ad hoc Wilson interval should be used as interval estimator

The ad hoc Wilson interval outperforms the Wald interval estimator in all situations studied here. If the Wilson interval estimator is used credible estimates can be produced for all domains with $n > 100$ as well as for all domains when it comes to the variable "Any crime against persons". This would considerable improve the number of regional estimates from the SCS. But to be able to produce intervals for the more rare crime types it is more crucial that variable setup 4 is used.

3.4.3. Disseminating statistics on different crime types can be done – but is it interesting?

This study shows that it is possible to produce interval estimates even for small domains. The question is if it is interesting to disseminate intervals that are very wide. Comparisons are not possible and the estimates as such may be too wide to be of any practical use. In the smallest domain the interval can be as wide as 30 percent. The question is then if it is interesting to disseminate such estimates,

especially if the point estimates of the domains are negative at the same time.

3.4.4. Point estimates or interval estimates

Today both point estimates and interval estimates are produced for county estimates of the variable "Any crime against persons". In this situation the confidence interval is calculated together with the point estimate but as domains get smaller the interval estimates get wider and at some point the question arises if the point estimates should be disseminated at all for some domains. When the intervals are sufficiently wide the point estimates are not very probable as an estimate at all.

3.4.5. Including municipalities in the sampling plan

If it is interesting to estimate smaller domains, such as domains in the size less than 100 the probability that the domain will not have a large enough sample in all strata is quite substantial. The only way to see to that domain estimation will be possible for all domains would be to include the municipalities in the sampling plan. But even then the non-response can introduce the situation that the variance cannot be estimated.

3.6. Further research

This study has answered some questions but at the same time additional question have arisen on how to improve and develop small area estimates regarding victimization and other rare binomial proportions in sample surveys.

One of the results of the study is the conclusion that a by adding the auxiliary information on whether the person has reported the crime or not the estimates are improved considerably. The question then arises if there are other variables that can be used from registers that are highly correlated with victimizations.

A second area of research is to find ways to reduce the proportion of negative GREG estimates for smaller proportions and areas.

A third area of research is to theoretically derive a Wilson interval for sample surveys or another similar interval based on the hypergeometric distribution.

The last area of research would be to use a different approach by using the Bayesian or the model-based framework to produce small area estimates in victimization surveys.

4. References

Brown, L. D., Cai, T. T. and DasGupta, A. (2001) "Interval estimation for a binomial proportion", *Statistical Science*, Vol, 16 No 2. pp. 101-117.

Brottsförebyggande rådet, Brå (2008). *Brottsoffers benägenhet att anmäla brott*. Author: Westfelt, L. Rapport 2008:12. Stockholm: Brottsförebyggande rådet.

Buelens, B. and Benchop, T. (2009) *Small area estimation of violent crime victim rates in the Netherlands*. Available online [2010-12-07]: http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/S1P1_SMALL_AREA_ESTIMATION_BUELENS_BENSCHOP.pdf

Estevao, V. M. and Särndal, C.-E. (2004) "Borrowing Strength Is Not the Best Technique Within a Wide Class of Design-Consistent Domain Estimators", *Journal of Official Statistics*, Vol.20, No.4, 2004. pp. 645–669.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2004) *Survey methodology*, Wiley: Hoboken, New Jersey.

Holt, D. and Smith, T.M.F. (1979) "Post Stratification", *Journal of the Royal Statistical Society. Series A (General)*, Vol. 142, No. 1, pp. 33-46.

Irlander, Å. and Westfelt, L. (2010:1) *NTU 2009: om utsatthet, trygghet och förtroende*. Stockholm: Brottsförebyggande rådet (BRÅ) Available online [2010-12-07]: http://www.bra.se/extra/measurepoint/?module_instance=4&name=2010_02_NTU_2009__rsbok.pdf&url=/dynamaster/file_archive/100128/c29984ddb3ae12ba92d16eed947143d8/2010%255f02%255fNTU%255f2009%255f%255frsbok.pdf

Irlander, Å. and Westfelt, L. (2010:2) *Nationella trygghetsundersökningen 2009. Teknisk rapport*. Stockholm: Brottsförebyggande rådet (BRÅ) Available online [2010-12-07]:: http://www.bra.se/extra/measurepoint/?module_instance=4&name=2010_02_NTU_2009__rsbok.pdf&url=/dynamaster/file_archive/100128/c29984ddb3ae12ba92d16eed947143d8/2010%255f02%255fNTU%255f2009%255f%255frsbok.pdf

Irlander, Å. and Wigerholt, J. (2009) *Nationella trygghetsundersökningen 2008. Teknisk rapport*. Stockholm: Brottsförebyggande rådet (BRÅ) Available online [2010-12-07]:: http://www.bra.se/extra/measurepoint/?module_instance=4&name=NTU_2008_teknisk_rapport.pdf&url=/dynamaster/file_archive/090428/1e4e1de22a42ed924198b9ee352e387c/NTU%255f2008%255fteknisk%255frapport.pdf

Korn, E. L. and Graubard, B. I. (1998) "Confidence intervals for proportions with small expected number of positive counts estimated from survey data", *Survey methodology*, vol 24, no 2, pp. 193-201.

Lehtonen, R. (2009) "Estimation for domains and small areas with designbased and model-based methods". Presentation at B-N-U Summer School, 23-27 August 2009, Kyiv, Ukraine. Available online [2010-12-07]: http://probability.univ.kiev.ua/school09/papers/Summer_School_Lehtonen_Slides.pdf

Lehtonen, R. and Pahkinen, E. J. (2004) *Practical methods for design and analysis of complex surveys*. 2. ed. Chichester: Wiley

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005) "Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains" in *Statistics in Transition*, 7, 649–673.

Lehtonen, R. and Veijanen, A. (2009) "Design-based methods of estimation for domains and small areas" in *Handbook of statistics, vol. 29:B, Sample surveys: inference and analysis*, Amsterdam Elsevier: North Holland.

Lundström, S. and Särndal, C.-E. (2001) *Estimation in the presence of nonresponse and frame Imperfections*, Statistiska Centralbyrån: Örebro.

Myrskylä, M. (2007) *Generalised regression estimation for domain class frequencies*, Research report no 247, Statistics Finland: Helsinki.

Ott, P. K. (2007) "Comparing design-based and model-based inference: an Introduction", *Biometrics Information Pamphlet 63*. B.C. Min. For. Range, Res. Br. Victoria, B.C.

Rao, J. N. K. (2003). *Small area estimation*. Wiley-Interscience, Hoboken, N.J.

Stukel, D.M., Hidiroglou, M.A., Särndal, C.-E. (1996) "Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization" *Survey Methodology*, Vol. 22, No 2, pp. 117-125.

Särndal, C.-E., Swensson, B. and Wretman, J. H. (1992) *Model assisted survey sampling*. New York: Springer-Vlg

Thorburn, D. (2009) "Bayesian Methods in Survey Sampling". Preliminary Version, Workshop on Survey Sampling, August 23-27, 2009, Kyiv.

Töyrä, A. (2007). *Nationella trygghetsundersökningen 2006. Teknisk rapport*. Stockholm: Brottsförebyggande rådet (BRÅ) Available online [2010-12-07]: http://www.bra.se/extra/measurepoint/?module_instance=4&name=01NTU_teknisk_slutred.pdf&url=/dynamaster/file_archive/070830/085e85aa25e32f3fba93e39aa6f3ca86/01NTU%255fteknisk%255fslutred.pdf

Töyrä, A. and Wigerholt, J. (2008). *Nationella trygghetsundersökningen 2007. Teknisk rapport*. Stockholm: Brottsförebyggande rådet (BRÅ) Availible online [2010-12-07]: http://www.bra.se/extra/measurepoint/?module_instance=4&name=2008_5_Nationella_trygghetsunders%f6kningen_2007_Teknisk_rapport.pdf&url=/dynamaster/file_archive/080521/f1631b6fe29bdf9e389d357945e1a132/2008%255f5%255fNationella%255ftrygghetsunders%25f6kningen%255f2007%255fTeknisk%255frapport.pdf

Appendix 1. Clustering of counties and municipalities

The counties were clustered together into seven different counties of different size and the municipalities were clustered together into 97 different artificial municipalities.

CODE	COUNTY	NAME
1	1	Stockholms län
3	1	Uppsala län
4	1	Södermanlands län
5	1	Östergötlands län
6	2	Jönköpings län
7	2	Kronobergs län
8	3	Kalmar län
9	3	Gotlands län
10	3	Blekinge län
12	4	Skåne län
13	4	Hallands län
14	4	Västra Götalands län
17	4	Värmlands län
18	5	Örebro län
19	5	Västmanlands län
20	6	Dalarnas län
21	6	Gävleborgs län
22	6	Västernorrlands län
23	7	Jämtlands län
24	7	Västerbottens län
25	7	Norrbottens län

CODE	MUNICIPALITY	NAME
0114	1	Upplands-Väsby
0115	1	Vallentuna
0117	1	Österåker
0120	2	Värmdö
0123	2	Järfälla
0125	2	Ekerö
0126	3	Huddinge
0127	3	Botkyrka
0128	3	Salem
0136	4	Haninge
0138	4	Tyresö
0139	4	Upplands-Bro
0140	5	Nykvarn
0160	5	Täby
0162	5	Danderyd
0163	6	Sollentuna
0180	6	Stockholm
0181	6	Södertälje
0182	7	Nacka

0183	7	Sundbyberg
0184	7	Solna
0186	8	Lidingö
0187	8	Vaxholm
0188	8	Norrtälje
0191	9	Sigtuna
0192	9	Nynäshamn
0305	9	Håbo
0319	10	Älvkarleby
0330	10	Knivsta
0331	10	Heby
0360	11	Tierp
0380	11	Uppsala
0381	11	Enköping
0382	12	Östhammar
0428	12	Vingåker
0461	12	Gnesta
0480	13	Nyköping
0481	13	Oxelösund
0482	13	Flen
0483	14	Katrineholm
0484	14	Eskilstuna
0486	14	Strängnäs
0488	15	Trosa
0509	15	Ödeshög
0512	15	Ydre
0513	16	Kinda
0560	16	Boxholm
0561	16	Åtvidaberg
0562	17	Finspång
0563	17	Valdemarsvik
0580	17	Linköping
0581	18	Norrköping
0582	18	Söderköping
0583	18	Motala
0584	19	Vadstena
0586	19	Mjölby
0604	19	Aneby
0617	20	Gnosjö
0642	20	Mullsjö
0643	20	Habo
0662	21	Gislaved
0665	21	Vaggeryd
0680	21	Jönköping
0682	22	Nässjö
0683	22	Värnamo
0684	22	Sävsjö
0685	23	Vetlanda
0686	23	Eksjö
0687	23	Tranås
0760	24	Uppvidinge

0761	24	Lessebo
0763	24	Tingsryd
0764	25	Alvesta
0765	25	Älmhult
0767	25	Markaryd
0780	26	Växjö
0781	26	Ljungby
0821	26	Högsby
0834	27	Torsås
0840	27	Mörbylånga
0860	27	Hultsfred
0861	28	Mönsterås
0862	28	Emmaboda
0880	28	Kalmar
0881	29	Nybro
0882	29	Oskarshamn
0883	29	Västervik
0884	30	Vimmerby
0885	30	Borgholm
0980	30	Gotland
1060	31	Olofström
1080	31	Karlskrona
1081	31	Ronneby
1082	32	Karlshamn
1083	32	Sölvesborg
1214	32	Svalöv
1230	33	Staffanstorp
1231	33	Burlöv
1233	33	Vellinge
1256	34	Östra Göinge
1257	34	Örkelljunga
1260	34	Bjuv
1261	35	Kävlinge
1262	35	Lomma
1263	35	Svedala
1264	36	Skurup
1265	36	Sjöbo
1266	36	Hörby
1267	37	Höör
1270	37	Tomelilla
1272	37	Bromölla
1273	38	Osby
1275	38	Perstorp
1276	38	Klippan
1277	39	Åstorp
1278	39	Båstad
1280	39	Malmö
1281	40	Lund
1282	40	Landskrona
1283	40	Helsingborg
1284	41	Höganäs

1285	41	Eslöv
1286	41	Ystad
1287	42	Trelleborg
1290	42	Kristianstad
1291	42	Simrishamn
1292	43	Ängelholm
1293	43	Hässleholm
1315	43	Hylte
1380	44	Halmstad
1381	44	Laholm
1382	44	Falkenberg
1383	45	Varberg
1384	45	Kungsbacka
1401	45	Härryda
1402	46	Partille
1407	46	Öckerö
1415	46	Stenungsund
1419	47	Tjörn
1421	47	Orust
1427	47	Sotenäs
1430	48	Munkedal
1435	48	Tanum
1438	48	Dals-Ed
1439	49	Färgelanda
1440	49	Ale
1441	49	Lerum
1442	50	Vårgårda
1443	50	Bollebygd
1444	50	Grästorp
1445	51	Essunga
1446	51	Karlsborg
1447	51	Gullspång
1452	52	Tranemo
1460	52	Bengtsfors
1461	52	Mellerud
1462	53	Lilla Edet
1463	53	Mark
1465	53	Svenljunga
1466	54	Herrljunga
1470	54	Vara
1471	54	Götene
1472	55	Tibro
1473	55	Töreboda
1480	55	Göteborg
1481	56	Mölndal
1482	56	Kungälv
1484	56	Lysekil
1485	57	Uddevalla
1486	57	Strömstad
1487	57	Vänersborg
1488	58	Trollhättan

1489	58	Alingsås
1490	58	Borås
1491	59	Ulricehamn
1492	59	Åmål
1493	59	Mariestad
1494	60	Lidköping
1495	60	Skara
1496	60	Skövde
1497	61	Hjo
1498	61	Tidaholm
1499	61	Falköping
1715	62	Kil
1730	62	Eda
1737	62	Torsby
1760	63	Storfors
1761	63	Hammarö
1762	63	Munkfors
1763	64	Forshaga
1764	64	Grums
1765	64	Årjäng
1766	65	Sunne
1780	65	Karlstad
1781	65	Kristinehamn
1782	66	Filipstad
1783	66	Hagfors
1784	66	Arvika
1785	67	Säffle
1814	67	Lekeberg
1860	67	Laxå
1861	68	Hallsberg
1862	68	Degerfors
1863	68	Hällefors
1864	69	Ljusnarsberg
1880	69	Örebro
1881	69	Kumla
1882	70	Askersund
1883	70	Karlskoga
1884	70	Nora
1885	71	Lindesberg
1904	71	Skinnskatteberg
1907	71	Surahammar
1917	72	Heby
1960	72	Kungsör
1961	72	Hallstahammar
1962	73	Norberg
1980	73	Västerås
1981	73	Sala
1982	74	Fagersta
1983	74	Köping
1984	74	Arboga
2021	75	Vansbro

2023	75	Malung
2026	75	Gagnef
2029	76	Leksand
2031	76	Rättvik
2034	76	Orsa
2039	77	Älvdalen
2061	77	Smedjebacken
2062	77	Mora
2080	78	Falun
2081	78	Borlänge
2082	78	Säter
2083	79	Hedemora
2084	79	Avesta
2085	79	Ludvika
2101	80	Ockelbo
2104	80	Hofors
2121	80	Ovanåker
2132	81	Nordanstig
2161	81	Ljusdal
2180	81	Gävle
2181	82	Sandviken
2182	82	Söderhamn
2183	82	Bollnäs
2184	83	Hudiksvall
2260	83	Ånge
2262	83	Timrå
2280	84	Härnösand
2281	84	Sundsvall
2282	84	Kramfors
2283	85	Sollefteå
2284	85	Örnsköldsvik
2303	85	Ragunda
2305	86	Bräcke
2309	86	Krokom
2313	86	Strömsund
2321	87	Åre
2326	87	Berg
2361	87	Härjedalen
2380	88	Östersund
2401	88	Nordmaling
2403	88	Bjurholm
2404	89	Vindeln
2409	89	Robertsfors
2417	89	Norsjö
2418	90	Malå
2421	90	Storuman
2422	90	Sorsele
2425	91	Dorotea
2460	91	Vännäs
2462	91	Vilhelmina
2463	92	Åsele

2480	92	Umeå
2481	92	Lycksele
2482	93	Skellefteå
2505	93	Arvidsjuar
2506	93	Arjeplog
2510	94	Jokkmokk
2513	94	Överkalix
2514	94	Kalix
2518	95	Övertorneå
2521	95	Pajala
2523	95	Gällivare
2560	96	Älvsbyn
2580	96	Luleå
2581	96	Piteå
2582	97	Boden
2583	97	Haparanda
2584	97	Kiruna

Appendix 2. Simulation program code (SAS)

The simulation made in SAS was done by running the program MASTER16.sas that includes the other SAS programs and the ANALYZE9.sas for the analysis of the simulation results.

MASTER16.sas

```
/* ========== LIBNAMES ========== */
libname DATA 'H:\Måns\SAE\DATA';
libname RESULTAT 'H:\Måns\SAE\RESULTAT';

/* ============================== */
/* PARAMETERS IN SIMULATION STUDY */
/* ============================== */

/* SIMULATION STARTS AT */
%let simu_start = 1;
/* NUMBER OF SIMULATIONS */
%let simu = 1000;
/* THE CURRENT SIMULATION */
%let simu_nr = &simu_start;

/* WHICH VARIABLE INDICATES STRATA */
%let INDIKATOR_STRATA = COUNTY;

/* INDICATES WHICH VARIABLES THAT ARE CLASS VARIABLES */
%let CLASSVAR = CIVIL_STATUS NON_NORDIC SEX MUNICIPALITY COUNTY URBAN AGECLASS INCOMECLASS;

/* SETUPS OF INDEPENDENT VARIABLES IN THE GREG-MODELS */
%LET INDEPVAR = SEX AGECLASS NON_NORDIC CIVIL_STATUS INCOMECLASS URBAN;
%LET DEPENDENT1 = COUNTY SEX AGECLASS NON_NORDIC CIVIL_STATUS INCOMECLASS URBAN;
%LET DEPENDENT2 = COUNTY SEX AGECLASS NON_NORDIC CIVIL_STATUS INCOMECLASS URBAN;
%LET DEPENDENT3 = MUNICIPALITY SEX AGECLASS NON_NORDIC CIVIL_STATUS INCOMECLASS URBAN;
%LET DEPENDENT4 = COUNTY SEX AGECLASS NON_NORDIC CIVIL_STATUS INCOMECLASS URBAN;

/* THE SEVEN TYPES OF Y-VARIABLES */
%LET RESPONSE1 = ROBBERY;
%LET RESPONSE2 = ASSAULT;
%LET RESPONSE3 = SEVERE_ASSAULT;
%LET RESPONSE4 = SEXUAL;
%LET RESPONSE5 = THREAT;
%LET RESPONSE6 = FRAUD;
%LET RESPONSE7 = HARASS;
%LET RESPONSE8 = ANY;

/* INDICATES THE NUMBER OF SETUPS OF INDEPENDENT VARIABLES (STARTS AT 1) AND Y-VARIABLES */
%let DEPNR = 4;
%let RESPONSENR = 8;
%LET DEPENDENTNR = &DEPNR;
%let LEVELNR = 2; /* MUNICIPALITY AND COUNTY */
%let LINKNR = 2; /* MUNICIPALITY AND COUNTY */
%let VARIABLENR = 2;

/* NUMBER OF MODELS USED IN GREG ESTIMATION */
%LET MODELNR = %eval(&DEPNR * &RESPONSENR * 2);
%LET TOT_MODELS = %eval(&DEPNR * &RESPONSENR * 2);

/* POPULATIONS SIZE, LARGE AND SMALL N*/
%LET POPSIZE = 51173;
%LET SAMPLESIZE = 5000;

/* INDICATES WHETHER THE LOG SHOULD BE CLEARED (LARGE SIMULATIONS) 1 = YES */
%LET CLEARLOG=0;

/* ============================== */
/* =========== STUDY ============ */
/* ============================== */


/* CREATING POPULATION */
%include 'H:\Måns\SAE\KOD\POPULATION9.sas';
```

```
/* CREATING DATASET TO STORE RESULTS FROM THE STUDY */
%include 'H:\Måns\SAE\KOD\STARTSETS5.sas';

/* READING MACROS FOR THE SIMULATION STUDY */
%include 'H:\Måns\SAE\KOD\MACROS33.sas';

%SIMULATION
```

MACRO33.sas

```
%MACRO SIMULATION;
/* ========== SIMULATION START ============ */
%DO k_sim=&simu_start %TO &simu;
              %let starttid=%sysfunc(datetime(),best.);
              %let simu_nr = &k_sim;

              /* DRAWING SAMPLES */
              %include 'H:\Måns\SAE\KOD\SAMPLE3.sas';

              /* FITTING MODELS  */
              %let MODELNR = 1;
              %MODELFIT

              /* ESTIMATION  */
              %LET LEVEL = COUNTY;
              %LET STRATA = COUNT_STR;
              %LET TOT_STRATA=21;
              %ESTIMATE_HT
              %ESTIMATE_GREG
              %ESTIMATE_ALL
              %LET LEVEL = MUNICIPALITY;
              %LET STRATA = MUNICIP_STR;
              %LET TOT_STRATA=291;
              %ESTIMATE_HT
              %ESTIMATE_GREG
              %ESTIMATE_ALL

              /* SAVING THE RESULTS  */
              %SAVE_EST_RESULTS

              /* CLEANING LOG WINDOW  */
              %IF &CLEARLOG=1 %THEN %DO;
              DM LOG 'CLEAR';
              %END;

              /* MEASURING SIMULATION TIME */
              %let sluttid=%sysfunc(datetime(),best.);
              DATA SIMULINFO;
                            SET SIMULINFO_BLANK;
                            SIMULERING=&simu_nr;
                            STARTTID=&starttid;
                            SLUTTID=&sluttid;
                            SIMULERINGSTID=&sluttid - &starttid;
              RUN;
              DATA RESULTAT.SIMULINFO;
                            SET RESULTAT.SIMULINFO SIMULINFO;
              RUN;

%END;

/* ADDING TRUE VALUES TO ESTIMATES - AFTER SIMULATION */
PROC SORT DATA=RESULTAT.ESTIMATES_COUNTY_&simu_start._&simu;
              BY COUNTY;
PROC SORT DATA=RESULTAT.ESTIMATES_MUNICIP_&simu_start._&simu;
              BY MUNICIPALITY;
RUN;

DATA RESULTAT.ESTIMATES_COUNTY_&simu_start._&simu;
              MERGE
              True_tot_values_county
              RESULTAT.ESTIMATES_COUNTY_&simu_start._&simu;
              BY COUNTY;
              IF SMALL_d_str_N = . THEN DELETE;
RUN;
DATA RESULTAT.ESTIMATES_MUNICIP_&simu_start._&simu;
```

```
            MERGE
            True_tot_values_MUNICIPALITY
            RESULTAT.ESTIMATES_MUNICIP_&simu_start._&simu;
            BY MUNICIPALITY;
            IF SMALL_d_str_N = . THEN DELETE;
RUN;

/* =========== SIMULATION END ============ */

%MEND;
%MACRO MODELFIT; /* FITTING MODELS TO THE SAMPLED DATA */
%DO i=1 %TO &RESPONSENR;
            %DO j=1 %TO &DEPNR;
                        %IF &i=1 %THEN %LET RESPONSE = &RESPONSE1;
                        %IF &i=2 %THEN %LET RESPONSE = &RESPONSE2;
                        %IF &i=3 %THEN %LET RESPONSE = &RESPONSE3;
                        %IF &i=4 %THEN %LET RESPONSE = &RESPONSE4;
                        %IF &i=5 %THEN %LET RESPONSE = &RESPONSE5;
                        %IF &i=6 %THEN %LET RESPONSE = &RESPONSE6;
                        %IF &i=7 %THEN %LET RESPONSE = &RESPONSE7;
                        %IF &i=8 %THEN %LET RESPONSE = &RESPONSE8;

                        %LET REPORTED = &RESPONSE;
                        %IF &RESPONSE = SEVERE_ASSAULT %THEN %LET REPORTED = ASSAULT;

                        %IF &j=1 %THEN %LET DEPENDENT = &DEPENDENT1;
                        %IF &j=2 %THEN %LET DEPENDENT = &DEPENDENT2
&REPORTED._REP_MUNICIPALITY;
                        %IF &j=3 %THEN %LET DEPENDENT = &DEPENDENT3;
                        %IF &j=4 %THEN %LET DEPENDENT = &DEPENDENT4 &REPORTED._REPORT;

                        ods exclude /* EXLUDE ALL OUTPUT */
                        ModelInfo
                        VarianceEstimation
                        NObs
                        ResponseProfile
                        ClassLevelInfo
                        ConvergenceStatus
                        FitStatistics
                        GlobalTests
                        Type3
                        ParameterEstimates
                        OddsRatios
                        Association
                        ;

                        PROC SURVEYLOGISTIC DATA=SIMULPOP_MODELFIT;
                                    STRATA STRATA_IND;
                                    ODS OUTPUT ConvergenceStatus=CONV_TEMP;
                                    OUTPUT OUT=TEMP1 PREDICTED=PRED_&MODELNR;
                                    CLASS &CLASSVAR;
                                    model &RESPONSE (DESCENDING) = &DEPENDENT;
                                    weight SamplingWeight;
                        RUN;
                        DATA SIMULPOP_MODELFIT;
                                    set TEMP1 (DROP = _LEVEL_);
                                    if SamplingIndicator = 1 then
                                    RES_&MODELNR = &RESPONSE - PRED_&MODELNR;
                        RUN;
                        DATA CONV_TEMP;
                                    set CONV_TEMP;
                                    MODEL = &MODELNR;
                                    SIMULATION = &simu_nr;
                        RUN;
                        DATA RESULTAT.CONVERGENCE;
                                    set RESULTAT.CONVERGENCE CONV_TEMP;
                        RUN;

                        %IF &simu_nr=1 %THEN %DO;
                        DATA MODELS2;
                                    set MODELS1;
                                    MODELNR="&MODELNR";
                                    TYPE='LOG';
                                    RESPONSE="&RESPONSE";
                                    DEPENDENT="&DEPENDENT";
                        RUN;
                        DATA RESULTAT.MODELS;
```

```
                                        set RESULTAT.MODELS MODELS2;
                                        if MODELNR="0" then DELETE;
                            RUN;
                            %END;

                            %let MODELNR = %eval(&MODELNR + 1);

                            ods exclude /* EXLUDE ALL OUTPUT */
                            DataSummary
                            DesignSummary
                            FitStatistics
                            ClassVarInfo
                            Effects
                            ;

                            PROC SURVEYREG DATA=SIMULPOP_MODELFIT;
                                        STRATA STRATA_IND;
                                        OUTPUT OUT=TEMP1 PREDICTED=PRED_&MODELNR;
                                        CLASS &CLASSVAR;
                                        model &RESPONSE = &DEPENDENT;
                                        weight SamplingWeight;
                            RUN;
                            DATA SIMULPOP_MODELFIT;
                                        set TEMP1;
                                        if SamplingIndicator = 1 then
                                        RES_&MODELNR = &RESPONSE - PRED_&MODELNR;
                            RUN;

                            %IF &simu_nr=1 %THEN %DO;
                            DATA MODELS2;
                                        set MODELS1;
                                        MODELNR="&MODELNR";
                                        TYPE='LIN';
                                        RESPONSE="&RESPONSE";
                                        DEPENDENT="&DEPENDENT";
                            RUN;
                            DATA RESULTAT.MODELS;
                                        set RESULTAT.MODELS MODELS2;
                                        if MODELNR="0" then DELETE;
                            RUN;
                            %END;

                            %let MODELNR = %eval(&MODELNR + 1);
                %END;
%END;
%MEND;
%MACRO ESTIMATE_HT;
PROC SORT DATA=SIMULPOP_MODELFIT OUT=SPOP_MF_&LEVEL;
                BY &STRATA;
RUN;

DATA SSAMP_MF_&LEVEL;
                set SPOP_MF_&LEVEL;
                if SamplingIndicator=1;
RUN;

PROC MEANS DATA=SSAMP_MF_&LEVEL NOPRINT; /* WEIGHTED SAMPLE */
                BY &STRATA;
                VAR &LEVEL ROBBERY -- ANY;
                OUTPUT
                OUT = EST_HT_2_&LEVEL
                MEAN(&LEVEL) = &LEVEL
                %DO j=1 %TO &RESPONSENR;
                            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
                            SUM(&RESPONSE) = HT_TOT_&RESPONSE
                %END;
                ;
                WEIGHT SamplingWeight;
RUN;
PROC MEANS DATA=SSAMP_MF_&LEVEL NOPRINT VARDEF=DF; /* SAMPLE WITHOUT WEIGHTS*/
```

```
              BY &STRATA;
              VAR &LEVEL ROBBERY -- ANY;
              OUTPUT
              OUT = EST_HT_3_&LEVEL
              MEAN(&LEVEL) = &LEVEL
              %DO j=1 %TO &RESPONSENR;
                            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
                            VAR(&RESPONSE) = HT_S2_&RESPONSE
              %END;
              ;
RUN;
DATA Est_ht_2_&LEVEL (DROP = _FREQ_ _TYPE_);
              SET Est_ht_2_&LEVEL;
              SMALL_d_N = _FREQ_;
RUN;
%MEND;
%MACRO ESTIMATE_GREG;
/* WHOLE POPULATION */
/* CALC SUM_(over U_d) y_hat */
PROC MEANS DATA=SPOP_MF_&LEVEL NOPRINT;
              BY &STRATA;
              VAR &LEVEL PRED_1 -- PRED_&TOT_MODELS;
              OUTPUT
              OUT = EST_GREG_1_&LEVEL
              MEAN(&LEVEL) = &LEVEL
              %DO i=1 %TO &TOT_MODELS;
                            SUM(PRED_&i) = TOT_YHAT_GREG&i
              %END;
              ;
RUN;

/* WEIGHTED SAMPLE */
/* CALC SUM_(over s_d) w*e - Summation of all weighted residuals */
PROC MEANS DATA=SSAMP_MF_&LEVEL NOPRINT;
              BY &STRATA;
              VAR &LEVEL RES_1 -- RES_&TOT_MODELS;
              OUTPUT
              OUT = EST_GREG_2_&LEVEL
              MEAN(&LEVEL) = &LEVEL
              %DO i=1 %TO &TOT_MODELS;
                            SUM(RES_&i) = TOT_WRES_GREG&i
              %END;
              ;
              WEIGHT SamplingWeight;
RUN;

/* CALC
Sum (e_dk - e_bar_d)^2
GREGi_SUM_SQ_RES
*/
%EDK

/* SAMPLE WITHOUT WEIGHTS*/
/* CALC E_S2_VAR(RES_i) = Sum((E_k - E_bar)^2/(n_d-1)) */
PROC MEANS DATA=SSAMP_MF_&LEVEL NOPRINT VARDEF=DF;
              BY &STRATA;
              VAR &LEVEL RES_1 -- RES_&TOT_MODELS;
              OUTPUT
              OUT = EST_GREG_4_&LEVEL
              MEAN(&LEVEL) = &LEVEL
              %DO i=1 %TO &TOT_MODELS;
                            VAR(RES_&i) = E_S2_GREG&i
              %END;
              ;
RUN;

%MEND;
%MACRO EDK;
/* SAMPLE WITHOUT WEIGHTS*/
/* CALC e_bar for all strata */
```

```
PROC MEANS DATA=SSAMP_MF_&LEVEL NOPRINT VARDEF=DF;
              BY &STRATA;
              VAR &LEVEL RES_1 -- RES_&TOT_MODELS;
              OUTPUT
              OUT = EST_E_BAR_&LEVEL
              MEAN(&LEVEL) = &LEVEL
              %DO i=1 %TO &TOT_MODELS;
                            MEAN(RES_&i) = E_bar_GREG&i
              %END;
              ;
RUN;

%DO j=1 %TO &TOT_STRATA;
DATA TEMP (DROP = SelectionProb -- RES_&TOT_MODELS);
              SET SSAMP_MF_&LEVEL (KEEP = &STRATA SelectionProb -- RES_&TOT_MODELS);
                            IF &STRATA = &j THEN DO;
                                          %DO i=1 %TO &TOT_MODELS;
                                          GREG_&i._E_dk = RES_&i;
                                          GREG_&i._E2_dk = (RES_&i)*(RES_&i);
                                          %END;
                                          END;
                            ELSE IF &STRATA ~= &j THEN DO;
                                          %DO i=1 %TO &TOT_MODELS;
                                          GREG_&i._E_dk = 0 - PRED_&i;
                                          GREG_&i._E2_dk = (0 - PRED_&i)*(0 - PRED_&i);
                                          %END;
                            END;
                            &STRATA=&j;
RUN;
PROC MEANS DATA=TEMP NOPRINT VARDEF=DF;
              VAR &STRATA GREG_1_E_dk -- GREG_&TOT_MODELS._E_dk;
              OUTPUT
              OUT = SUM_E_dk
                            MEAN(&STRATA) = &STRATA
                            %DO i=1 %TO &TOT_MODELS;
                                          SUM(GREG_&i._E_dk) = SUM_GREG_&i._E_dk
                                          SUM(GREG_&i._E2_dk) = SUM_GREG_&i._E2_dk
                            %END;
              ;
RUN;
%IF &j = 1 %THEN %DO;
DATA EDK_ALL_D;
              SET SUM_E_dk;
RUN;
%END;
%ELSE %IF &j > 1 %THEN %DO;
DATA EDK_ALL_D;
              SET EDK_ALL_D SUM_E_dk;
RUN;
%END;
%END;

/* CALC
Sum (e_dk - e_bar_d)^2
as
Sum (e_dk)^2 - 2 * e_bar_d * Sum(e_dk) + k*e_bar_d^2
See 6.61 in Lehtonen and Pahktinen (2004) p. 202
*/
DATA Est_GREG_3_&LEVEL
              (DROP = _TYPE_ _FREQ_ E_bar_GREG1 -- SUM_GREG_&TOT_MODELS._E2_dk)
              ;
              MERGE
              EST_E_BAR_&LEVEL
              EDK_ALL_D;
              BY &STRATA;
              if E_bar_GREG1 = . THEN DELETE;
              %DO i=1 %TO &TOT_MODELS;
                            GREG_&i._SUM_SQ_RES =
                            (SUM_GREG_&i._E2_dk) -
                            (2 * E_bar_GREG&i * SUM_GREG_&i._E_dk) +
                            (_FREQ_ * E_bar_GREG&i * E_bar_GREG&i)
                            ;
              %END;
RUN;
%MEND;
%MACRO ESTIMATE_ALL;
DATA ESTIMATES_&STRATA
```

```
                (DROP = _FREQ_
                TOT_YHAT_GREG1 -- TOT_YHAT_GREG&TOT_MODELS
                TOT_WRES_GREG1 -- TOT_WRES_GREG&TOT_MODELS
                GREG_1_SUM_SQ_RES -- GREG_&TOT_MODELS._SUM_SQ_RES
                HT_S2_&RESPONSE1 -- HT_S2_&RESPONSE8
                E_S2_GREG1 -- E_S2_GREG&TOT_MODELS
                );
                MERGE
                Est_GREG_1_&LEVEL (DROP = _TYPE_ )
                Est_GREG_2_&LEVEL (DROP = _TYPE_ _FREQ_)
                Est_GREG_3_&LEVEL
                Est_GREG_4_&LEVEL (DROP = _TYPE_ _FREQ_)
                Est_ht_2_&LEVEL
                Est_ht_3_&LEVEL (DROP = _TYPE_ _FREQ_);

                BY &STRATA;
                LARGE_d_N = _FREQ_;
                SMALL_N = %EVAL(&SAMPLESIZE);
                LARGE_N = %EVAL(&POPSIZE);


                /* DELETE STRATA WITH n<2 */
                if SMALL_d_N = 1 OR SMALL_d_N = . THEN DELETE;

                %DO j=1 %TO &RESPONSENR;
                                %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                                %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                                %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                                %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                                %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                                %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                                %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                                %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;

                                VAR_HT_&RESPONSE = LARGE_d_N*LARGE_d_N*(1-(SMALL_d_N/LARGE_d_N))*
(HT_S2_&RESPONSE / SMALL_d_N);
                %END;

                %DO i=1 %TO &TOT_MODELS;
                                GREG_&i = TOT_YHAT_GREG&i + TOT_WRES_GREG&i;
                %END;

                %DO i=1 %TO &TOT_MODELS;
                                /* CALCULATED AS 6.15 p. 202 in Lehtonen and Pahktinen (2004) */
                                VAR_P_GREG_&i = LARGE_d_N*LARGE_d_N*(1-
(SMALL_d_N/LARGE_d_N))*(1/SMALL_d_N)*E_S2_GREG&i;
                                /* CALCULATED AS 6.16 p. 202 in Lehtonen and Pahktinen (2004) */
                                VAR_U_GREG_&i = LARGE_N*LARGE_N*(1-(SMALL_N/LARGE_N))*(1/SMALL_N) *
(1/(SMALL_N-1))* GREG_&i._SUM_SQ_RES;
                %END;
RUN;

/* SUMMING UP VAR AND POINT ESTIMATES IN EACH STRATA IN EACH DOMAIN */
PROC SORT DATA=ESTIMATES_&STRATA;
                BY &LEVEL;
RUN;

PROC MEANS DATA=ESTIMATES_&STRATA NOPRINT;
                BY &LEVEL;
                VAR HT_TOT_&RESPONSE1 -- VAR_U_GREG_&TOT_MODELS;
                OUTPUT
                OUT = EST_TEMP_&LEVEL
                                SUM(SMALL_d_N) = SMALL_d_str_N
                                SUM(LARGE_d_N) = LARGE_d_str_N
                %DO j=1 %TO &RESPONSENR;
                                %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                                %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                                %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                                %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                                %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                                %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                                %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                                %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
                                SUM(HT_TOT_&RESPONSE) = HT_TOT_&RESPONSE
                                SUM(VAR_HT_&RESPONSE) = VAR_HT_&RESPONSE

                %END;
```

```
            %DO i=1 %TO &TOT_MODELS;
                        SUM(GREG_&i) = GREG_&i
                        SUM(VAR_P_GREG_&i) = VAR_P_GREG_&i
                        SUM(VAR_U_GREG_&i) = VAR_U_GREG_&i
            %END;
            ;
RUN;

DATA ESTIMATES_&LEVEL (DROP = _FREQ_ _TYPE_);
            SET EST_TEMP_&LEVEL;
            if _FREQ_ < 3 THEN DELETE;
RUN;

%MEND;
/*%MACRO ESTIMATE_RESULTS_HT;
DATA EST_TEMP_&LEVEL;
            SET EST_&LEVEL;
            %DO j=1 %TO &RESPONSENR;
                        %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                        %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                        %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                        %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                        %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                        %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                        %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                        %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;

                                    if TRUE_TOT_&RESPONSE < HT_TOT_&RESPONSE + &Z_alpha *
sqrt(VAR_PLANNED_HT_&RESPONSE)
                                    AND TRUE_TOT_&RESPONSE > HT_TOT_&RESPONSE - &Z_alpha
* sqrt(VAR_PLANNED_HT_&RESPONSE)
                                    then
                                    HT_INT_PL_&RESPONSE = 1;
                                    else
                                    HT_INT_PL_&RESPONSE = 0;

                                    if VAR_PLANNED_HT_&RESPONSE <= 0 then
                                    HT_PL_BRK_&RESPONSE = 1;
                                    else
                                    HT_PL_BRK_&RESPONSE = 0;

                                    if HT_TOT_&RESPONSE - &Z_alpha *
sqrt(VAR_PLANNED_HT_&RESPONSE) < 0
                                    then
                                    HT_INC_PL_&RESPONSE = 1;
                                    else
                                    HT_INC_PL_&RESPONSE = 0;
            %END;

RUN;
%MEND;*/
/*%MACRO ESTIMATE_RESULTS_GREG;
DATA ESTIMATES_&LEVEL;
            SET EST_TEMP_&LEVEL;
            %DO i=1 %TO &TOT_MODELS;
                        %IF &i>=1 AND &i <= %eval(&DEPENDENTNR*2*1)
                                                %THEN %LET RESPONSE =
&RESPONSE1;
                        %IF &i>=%eval(&DEPENDENTNR*2*1+1) AND &i<=%eval(&DEPENDENTNR*2*2)
            %THEN %LET RESPONSE = &RESPONSE2;
                        %IF &i>=%eval(&DEPENDENTNR*2*2+1) AND &i<=%eval(&DEPENDENTNR*2*3)
            %THEN %LET RESPONSE = &RESPONSE3;
                        %IF &i>=%eval(&DEPENDENTNR*2*3+1) AND &i<=%eval(&DEPENDENTNR*2*4)
            %THEN %LET RESPONSE = &RESPONSE4;
                        %IF &i>=%eval(&DEPENDENTNR*2*4+1) AND &i<=%eval(&DEPENDENTNR*2*5)
            %THEN %LET RESPONSE = &RESPONSE5;
                        %IF &i>=%eval(&DEPENDENTNR*2*5+1) AND &i<=%eval(&DEPENDENTNR*2*6)
            %THEN %LET RESPONSE = &RESPONSE6;
                        %IF &i>=%eval(&DEPENDENTNR*2*6+1) AND &i<=%eval(&DEPENDENTNR*2*7)
            %THEN %LET RESPONSE = &RESPONSE7;
                        %IF &i>=%eval(&DEPENDENTNR*2*7+1) AND &i<=%eval(&DEPENDENTNR*2*8)
            %THEN %LET RESPONSE = &RESPONSE8;

                                    if TRUE_TOT_&RESPONSE < GREG_&i + &Z_alpha *
sqrt(VAR_PLANNED_GREG_&i)
                                    AND TRUE_TOT_&RESPONSE > GREG_&i - &Z_alpha *
sqrt(VAR_PLANNED_GREG_&i)
```

```
                                        then
                                        GREG_INT_PL_&i = 1;
                                        else
                                        GREG_INT_PL_&i = 0;

                                        if VAR_PLANNED_GREG_&i <= 0 then
                                        GREG_PL_BRK_&i = 1;
                                        else
                                        GREG_PL_BRK_&i = 0;

                                        if GREG_&i < 0 then
                                        GREG_PL_NEG_&i = 1;
                                        else
                                        GREG_PL_NEG_&i = 0;

                                        if GREG_&i - &Z_alpha * sqrt(VAR_PLANNED_GREG_&i) < 0
                                        then
                                        GREG_INC_PL_&i = 1;
                                        else
                                        GREG_INC_PL_&i = 0;
                            %END;
                            SIMULATION = &simu_nr;
                RUN;
%MEND;*/
%MACRO SAVE_EST_RESULTS;
                %IF &simu_nr = &simu_start %THEN
                            %DO;
                            DATA RESULTAT.ESTIMATES_COUNTY_&simu_start._&simu;
                                        SET ESTIMATES_COUNTY;
                            RUN;
                            %END;
                %ELSE
                            %DO;
                            DATA RESULTAT.ESTIMATES_COUNTY_&simu_start._&simu;
                                        SET RESULTAT.ESTIMATES_COUNTY_&simu_start._&simu
ESTIMATES_COUNTY;
                            RUN;
                            %END;

                %IF &simu_nr = &simu_start %THEN
                            %DO;
                            DATA RESULTAT.ESTIMATES_MUNICIP_&simu_start._&simu;
                                        SET ESTIMATES_MUNICIPALITY;
                            RUN;
                            %END;
                %ELSE
                            %DO;
                            DATA RESULTAT.ESTIMATES_MUNICIP_&simu_start._&simu;
                                        SET RESULTAT.ESTIMATES_MUNICIP_&simu_start._&simu
ESTIMATES_MUNICIPALITY;
                            RUN;
                            %END;
%MEND;
%MACRO RRMSEARB;
DATA ESTIMATES_&LEVEL._ALL;
                set RESULTAT.ESTIMATES_&LEVEL._ALL;
                %DO j=1 %TO &RESPONSENR;
                            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;

                            DIFF_HT_TOT_&RESPONSE = HT_TOT_&RESPONSE - TRUE_TOT_&RESPONSE;
                            DIFFSQ_HT_TOT_&RESPONSE = (HT_TOT_&RESPONSE -
TRUE_TOT_&RESPONSE)**2;

                %END;

                %DO i=1 %TO &TOT_MODELS;
                            %IF &i>=1 AND &i <= %eval(&DEPENDENTNR*2*1)
                                                                    %THEN %LET RESPONSE =
&RESPONSE1;
```

```
                        %IF &i>=%eval(&DEPENDENTNR*2*1+1) AND &i<=%eval(&DEPENDENTNR*2*2)
            %THEN %LET RESPONSE = &RESPONSE2;
                        %IF &i>=%eval(&DEPENDENTNR*2*2+1) AND &i<=%eval(&DEPENDENTNR*2*3)
            %THEN %LET RESPONSE = &RESPONSE3;
                        %IF &i>=%eval(&DEPENDENTNR*2*3+1) AND &i<=%eval(&DEPENDENTNR*2*4)
            %THEN %LET RESPONSE = &RESPONSE4;
                        %IF &i>=%eval(&DEPENDENTNR*2*4+1) AND &i<=%eval(&DEPENDENTNR*2*5)
            %THEN %LET RESPONSE = &RESPONSE5;
                        %IF &i>=%eval(&DEPENDENTNR*2*5+1) AND &i<=%eval(&DEPENDENTNR*2*6)
            %THEN %LET RESPONSE = &RESPONSE6;
                        %IF &i>=%eval(&DEPENDENTNR*2*6+1) AND &i<=%eval(&DEPENDENTNR*2*7)
            %THEN %LET RESPONSE = &RESPONSE7;
                        %IF &i>=%eval(&DEPENDENTNR*2*7+1) AND &i<=%eval(&DEPENDENTNR*2*8)
            %THEN %LET RESPONSE = &RESPONSE8;

                        DIFF_GREG_&i = GREG_&i - TRUE_TOT_&RESPONSE;
                        DIFFSQ_GREG_&i = (GREG_&i - TRUE_TOT_&RESPONSE)**2;

            %END;

RUN;
%MEND;
%MACRO AGGREGATE;
PROC SORT DATA=ESTIMATES_&LEVEL._ALL;
            BY &LEVEL;
RUN;
PROC MEANS  NOPRINT DATA=ESTIMATES_&LEVEL._ALL;
            BY &LEVEL;
            VAR TRUE_TOT_ROBBERY -- WALD_Bra_INC_64;
            OUTPUT OUT=SUMMARY_&LEVEL
            MEAN(LARGE_N)=LARGE_N
            MEAN(SMALL_N)=SMALL_N_MEAN

            %DO j=1 %TO &RESPONSENR;
                        %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                        %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                        %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                        %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                        %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                        %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                        %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                        %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;

                        MEAN(TRUE_TOT_&RESPONSE) = TRUE_TOT_&RESPONSE

                        MEAN(HT_INT_PL_&RESPONSE) = HT_INT_PL_&RESPONSE

                        MEAN(HT_PL_BRK_&RESPONSE) = HT_PL_BRK_&RESPONSE
                        MEAN(HT_INC_PL_&RESPONSE) = HT_INC_PL_&RESPONSE

                        MEAN(DIFF_HT_TOT_&RESPONSE) = DIFF_HT_TOT_&RESPONSE
                        MEAN(DIFFSQ_HT_TOT_&RESPONSE) = DIFFSQ_HT_TOT_&RESPONSE
            %END;

            %DO i=1 %TO &TOT_MODELS;
                        MEAN(GREG_INT_PL_&i) = GREG_INT_PL_&i

                        MEAN(GREG_PL_BRK_&i) = GREG_PL_BRK_&i

                        MEAN(GREG_PL_NEG_&i) = GREG_PL_NEG_&i

                        MEAN(WILSON_BASIC_COV_&i) = WILSON_BASIC_COV_&i
                        MEAN(WALD_GREG_COV_&i) = WALD_GREG_COV_&i
                        MEAN(WALD_Bra_COV_&i) = WALD_Bra_COV_&i
                        MEAN(WILSON_WIDTH_BASIC_&i) = WILSON_WIDTH_MEAN_&i
                        MEAN(WALD_GREG_WIDTH_&i) = WALD_GREG_WIDTH_MEAN_&i
                        MEAN(WALD_Bra_WIDTH_&i) = WALD_Bra_WIDTH_MEAN_&i
                        MEAN(WILSON_BASIC_INC_&i) = WILSON_BASIC_INC_&i
                        MEAN(WALD_GREG_INC_&i) = WALD_GREG_INC_&i
                        MEAN(WALD_Bra_INC_&i) = WALD_Bra_INC_&i

                        MEAN(DIFF_GREG_&i) = DIFF_GREG_&i
                        MEAN(DIFFSQ_GREG_&i) = DIFFSQ_GREG_&i
            %END;
            ;
RUN;
DATA RESULTS_&LEVEL (DROP =
```

```
              %DO j=1 %TO &RESPONSENR;
                            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
                            DIFF_HT_TOT_&RESPONSE
                            DIFFSQ_HT_TOT_&RESPONSE
              %END;
              %DO i=1 %TO &TOT_MODELS;
                            DIFF_GREG_&i
                            DIFFSQ_GREG_&i
              %END;
);
              set SUMMARY_&LEVEL;

              %DO j=1 %TO &RESPONSENR;
                            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
                            if TRUE_TOT_&RESPONSE = 0 then TRUE_TOT_NOZERO_&RESPONSE = 1;
                            else TRUE_TOT_NOZERO_&RESPONSE = TRUE_TOT_&RESPONSE;
                            ARB_HT_TOT_&RESPONSE = abs(DIFF_HT_TOT_&RESPONSE) /
(TRUE_TOT_NOZERO_&RESPONSE);
                            RRMSE_HT_TOT_&RESPONSE = sqrt(DIFFSQ_HT_TOT_&RESPONSE) /
(TRUE_TOT_NOZERO_&RESPONSE);
              %END;

              %DO i=1 %TO &TOT_MODELS;
                            %IF &i>=1 AND &i <= %eval(&DEPENDENTNR*2*1)
                                                                        %THEN %LET RESPONSE =
&RESPONSE1;
                            %IF &i>=%eval(&DEPENDENTNR*2*1+1) AND &i<=%eval(&DEPENDENTNR*2*2)
              %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &i>=%eval(&DEPENDENTNR*2*2+1) AND &i<=%eval(&DEPENDENTNR*2*3)
              %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &i>=%eval(&DEPENDENTNR*2*3+1) AND &i<=%eval(&DEPENDENTNR*2*4)
              %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &i>=%eval(&DEPENDENTNR*2*4+1) AND &i<=%eval(&DEPENDENTNR*2*5)
              %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &i>=%eval(&DEPENDENTNR*2*5+1) AND &i<=%eval(&DEPENDENTNR*2*6)
              %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &i>=%eval(&DEPENDENTNR*2*6+1) AND &i<=%eval(&DEPENDENTNR*2*7)
              %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &i>=%eval(&DEPENDENTNR*2*7+1) AND &i<=%eval(&DEPENDENTNR*2*8)
              %THEN %LET RESPONSE = &RESPONSE8;
                            ARB_GREG_&i = abs(DIFF_GREG_&i) / TRUE_TOT_NOZERO_&RESPONSE;
                            RRMSE_GREG_&i = sqrt(DIFFSQ_GREG_&i) / TRUE_TOT_NOZERO_&RESPONSE;
              %END;
RUN;
%MEND;
%MACRO INCORRECT_VALUES;
PROC MEANS DATA=RESULTAT.RESULTS;
              CLASS DOMAINCLASS;
              VAR
              %DO i=1 %TO &TOT_MODELS;
              GREG_PL_NEG_&i
              %END;
              ;
RUN;
PROC MEANS DATA=RESULTAT.RESULTS;
              CLASS DOMAINCLASS;
              VAR
              %DO i=1 %TO &TOT_MODELS;
              GREG_INC_PL_&i
              %END;
              ;
RUN;
%MEND;
```

```
%MACRO INTERVAL_COVERAGE;
PROC MEANS DATA=RESULTAT.RESULTS;
              CLASS DOMAINCLASS;
              VAR
              %DO i=1 %TO &TOT_MODELS;
              GREG_INT_PL_&i
              %END;
              ;
RUN;
%MEND;
%MACRO ARB_AND_RRMSE;
PROC MEANS DATA=RESULTAT.RESULTS;
              CLASS DOMAINCLASS;
              VAR
              %DO i=1 %TO &TOT_MODELS;
              RRMSE_GREG_&i
              %END;
              ;
RUN;
PROC MEANS DATA=RESULTAT.RESULTS;
              CLASS DOMAINCLASS;
              VAR
              %DO i=1 %TO &TOT_MODELS;
              ARB_GREG_&i
              %END;
              ;
RUN;
DATA RESULT_TEMP;
   SET RESULTAT.RESULTS;
   IF RRMSE_HT_TOT_ROBBERY = 0 THEN RRMSE_HT_TOT_ROBBERY = . ;
   IF RRMSE_HT_TOT_ASSAULT = 0 THEN RRMSE_HT_TOT_ASSAULT = . ;
   IF RRMSE_HT_TOT_SEVERE_ASSAULT = 0 THEN RRMSE_HT_TOT_SEVERE_ASSAULT = . ;
   IF RRMSE_HT_TOT_SEXUAL = 0 THEN RRMSE_HT_TOT_SEXUAL = . ;
   IF RRMSE_HT_TOT_THREAT = 0 THEN RRMSE_HT_TOT_THREAT = . ;
   IF RRMSE_HT_TOT_FRAUD = 0 THEN RRMSE_HT_TOT_FRAUD = . ;
   IF RRMSE_HT_TOT_HARASS = 0 THEN RRMSE_HT_TOT_HARASS = . ;
   IF RRMSE_HT_TOT_ANY = 0 THEN RRMSE_HT_TOT_ANY = . ;
   IF ARB_HT_TOT_ROBBERY = 0 THEN ARB_HT_TOT_ROBBERY = . ;
   IF ARB_HT_TOT_ASSAULT = 0 THEN ARB_HT_TOT_ASSAULT = . ;
   IF ARB_HT_TOT_SEVERE_ASSAULT = 0 THEN ARB_HT_TOT_SEVERE_ASSAULT = . ;
   IF ARB_HT_TOT_SEXUAL = 0 THEN ARB_HT_TOT_SEXUAL = . ;
   IF ARB_HT_TOT_THREAT = 0 THEN ARB_HT_TOT_THREAT = . ;
   IF ARB_HT_TOT_FRAUD = 0 THEN ARB_HT_TOT_FRAUD = . ;
   IF ARB_HT_TOT_HARASS = 0 THEN ARB_HT_TOT_HARASS = . ;
   IF ARB_HT_TOT_ANY = 0 THEN ARB_HT_TOT_ANY = . ;
RUN;

PROC MEANS DATA=RESULT_TEMP;
              CLASS DOMAINCLASS;
              VAR
    RRMSE_HT_TOT_ROBBERY
    RRMSE_HT_TOT_ASSAULT
    RRMSE_HT_TOT_SEVERE_ASSAULT
    RRMSE_HT_TOT_SEXUAL
    RRMSE_HT_TOT_THREAT
    RRMSE_HT_TOT_FRAUD
    RRMSE_HT_TOT_HARASS
    RRMSE_HT_TOT_ANY
              ;
RUN;
PROC MEANS DATA=RESULT_TEMP;
              CLASS DOMAINCLASS;
              VAR
    ARB_HT_TOT_ROBBERY
    ARB_HT_TOT_ASSAULT
    ARB_HT_TOT_SEVERE_ASSAULT
    ARB_HT_TOT_SEXUAL
    ARB_HT_TOT_THREAT
    ARB_HT_TOT_FRAUD
    ARB_HT_TOT_HARASS
    ARB_HT_TOT_ANY
              ;
RUN;
%MEND;
%MACRO MODEL_CHECK; /* FITTING MODELS TO THE SAMPLED DATA */
%LET MODELNR = 1;
DATA TEMP_POP;
```

```
            SET DATA.POPULATION;
            W=1;
RUN;
DATA RESULTAT.Association;
RUN;
DATA RESULTAT.FitStatistics;
RUN;
%DO i=1 %TO &RESPONSENR;
            %DO j=1 %TO &DEPNR;
                        %IF &i=1 %THEN %LET RESPONSE = &RESPONSE1;
                        %IF &i=2 %THEN %LET RESPONSE = &RESPONSE2;
                        %IF &i=3 %THEN %LET RESPONSE = &RESPONSE3;
                        %IF &i=4 %THEN %LET RESPONSE = &RESPONSE4;
                        %IF &i=5 %THEN %LET RESPONSE = &RESPONSE5;
                        %IF &i=6 %THEN %LET RESPONSE = &RESPONSE6;
                        %IF &i=7 %THEN %LET RESPONSE = &RESPONSE7;
                        %IF &i=8 %THEN %LET RESPONSE = &RESPONSE8;

                        %LET REPORTED = &RESPONSE;
                        %IF &RESPONSE = SEVERE_ASSAULT %THEN %LET REPORTED = ASSAULT;

                        %IF &j=1 %THEN %LET DEPENDENT = &DEPENDENT1;
                        %IF &j=2 %THEN %LET DEPENDENT = &DEPENDENT2
&REPORTED._REP_MUNICIPALITY;
                        %IF &j=3 %THEN %LET DEPENDENT = &DEPENDENT3;
                        %IF &j=4 %THEN %LET DEPENDENT = &DEPENDENT4 &REPORTED._REPORT;

                        PROC SURVEYLOGISTIC DATA=TEMP_POP;
                                    CLASS &CLASSVAR;
                                    ODS OUTPUT Association=ASSO_TEMP;
                                    model &RESPONSE (DESCENDING) = &DEPENDENT;
                                    Weight W;
                        RUN;
                        DATA ASSO_TEMP;
                                    set ASSO_TEMP;
                                    MODEL = &MODELNR;
                        RUN;
                        DATA RESULTAT.Association;
                                    set RESULTAT.Association ASSO_TEMP;
                        RUN;

                        %let MODELNR = %eval(&MODELNR + 1);

                        PROC SURVEYREG DATA=TEMP_POP;
                                    CLASS &CLASSVAR;
                                    ODS OUTPUT FitStatistics=FIT_TEMP;
                                    model &RESPONSE = &DEPENDENT;
                                    Weight W;
                        RUN;
                        DATA FIT_TEMP;
                                    set FIT_TEMP;
                                    MODEL = &MODELNR;
                        RUN;
                        DATA RESULTAT.FitStatistics;
                                    set RESULTAT.FitStatistics FIT_TEMP;
                        RUN;

                        %let MODELNR = %eval(&MODELNR + 1);
            %END;
%END;
%MEND;
%MACRO INTERVAL_COVERAGE_2;
PROC MEANS DATA=RESULTAT.RESULTS;
            CLASS DOMAINCLASS;
            VAR
            %DO i=1 %TO &TOT_MODELS;
            WILSON_BASIC_COV_&i
            %END;
            %DO i=1 %TO &TOT_MODELS;
            WALD_GREG_COV_&i
            %END;
            %DO i=1 %TO &TOT_MODELS;
            WALD_Bra_COV_&i
            %END;
            ;
RUN;
%MEND;
```

```
%MACRO WILSON_COVERAGE_NOZERO;
%DO j=1 %TO 8;
              %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
              %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
              %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
              %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
              %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
              %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
              %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
              %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
                            PROC MEANS DATA=TEMP;
                                          CLASS &RESPONSE._ZERO;
                                          VAR
                                          %DO i=(((%EVAL(&j)-1)*8)+1)%TO (%EVAL(&j)*8);
                                          WILSON_BASIC_COV_&i
                                          %END;
                                          ;
                            RUN;
%END;
%MEND;


%MACRO INCORRECT_INTERVALS;
PROC MEANS DATA=RESULTAT.RESULTS;
              CLASS DOMAINCLASS;
              VAR
              %DO i=1 %TO &TOT_MODELS;
              WILSON_BASIC_INC_&i
              %END;
              %DO i=1 %TO &TOT_MODELS;
              WALD_GREG_INC_&i
              %END;
              %DO i=1 %TO &TOT_MODELS;
              WALD_Bra_INC_&i
              %END;
              ;
RUN;
%MEND;
%MACRO INTERVAL_WIDTH;
PROC MEANS DATA=RESULTAT.RESULTS;
              CLASS DOMAINCLASS;
              VAR
              %DO i=1 %TO &TOT_MODELS;
              WILSON_WIDTH_MEAN_&i
              %END;
              %DO i=1 %TO &TOT_MODELS;
              WALD_GREG_WIDTH_MEAN_&i
              %END;
              %DO i=1 %TO &TOT_MODELS;
              WALD_Bra_WIDTH_MEAN_&i
              %END;
              ;
RUN;
%MEND;
%MACRO INTERVAL_GREG;
DATA ESTIMATES_&LEVEL._ALL (DROP = var_p -- Wald_Bra_low);
              SET ESTIMATES_&LEVEL._ALL;
              %DO i=1 %TO &TOT_MODELS;
                            %IF &i>=1 AND &i <= %eval(&DEPENDENTNR*2*1)
                                                                                %THEN %LET RESPONSE =
&RESPONSE1;
                            %IF &i>=%eval(&DEPENDENTNR*2*1+1) AND &i<=%eval(&DEPENDENTNR*2*2)
              %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &i>=%eval(&DEPENDENTNR*2*2+1) AND &i<=%eval(&DEPENDENTNR*2*3)
              %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &i>=%eval(&DEPENDENTNR*2*3+1) AND &i<=%eval(&DEPENDENTNR*2*4)
              %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &i>=%eval(&DEPENDENTNR*2*4+1) AND &i<=%eval(&DEPENDENTNR*2*5)
              %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &i>=%eval(&DEPENDENTNR*2*5+1) AND &i<=%eval(&DEPENDENTNR*2*6)
              %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &i>=%eval(&DEPENDENTNR*2*6+1) AND &i<=%eval(&DEPENDENTNR*2*7)
              %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &i>=%eval(&DEPENDENTNR*2*7+1) AND &i<=%eval(&DEPENDENTNR*2*8)
              %THEN %LET RESPONSE = &RESPONSE8;

              var_p = VAR_PLANNED_GREG_&i/(LARGE_N*LARGE_N);
```

```
            z = &Z_alpha;
            n = SMALL_N;
            IF GREG_PL_NEG_&i = 0 THEN p = (GREG_&i / LARGE_N);
            IF GREG_PL_NEG_&i = 1 THEN p = 0;
            Wilson_BASIC_up = round((p + ((z*z)/(2*n)) + z*Sqrt((p*(1-p))/n +
(z*z)/(4*n*n)))/(1+(z*z)/n),.000001);
            Wilson_BASIC_low = round((p + ((z*z)/(2*n)) - z*Sqrt((p*(1-p))/n +
(z*z)/(4*n*n)))/(1+(z*z)/n),.000001);
            Wald_GREG_up= round(p + z * Sqrt(var_p),.000001);
            Wald_GREG_low= round(p - z * Sqrt(var_p),.000001);
            Wald_Bra_up= round(p + z * Sqrt(p*(1-p)/n),.000001);
            Wald_Bra_low= round(p - z * Sqrt(p*(1-p)/n),.000001);
            IF (TRUE_TOT_&RESPONSE / LARGE_N) <= Wilson_BASIC_up AND (TRUE_TOT_&RESPONSE /
LARGE_N) >= Wilson_BASIC_low
            THEN WILSON_BASIC_COV_&i = 1;
            ELSE WILSON_BASIC_COV_&i = 0;
            IF (TRUE_TOT_&RESPONSE / LARGE_N) <= Wald_GREG_up AND (TRUE_TOT_&RESPONSE /
LARGE_N) >= Wald_GREG_low
            THEN WALD_GREG_COV_&i = 1;
            ELSE WALD_GREG_COV_&i = 0;
            IF (TRUE_TOT_&RESPONSE / LARGE_N) <= Wald_Bra_up AND (TRUE_TOT_&RESPONSE /
LARGE_N) >= Wald_Bra_low
            THEN WALD_Bra_COV_&i = 1;
            ELSE WALD_Bra_COV_&i = 0;
            WILSON_WIDTH_BASIC_&i = Wilson_BASIC_up - Wilson_BASIC_low;
            WALD_GREG_WIDTH_&i = Wald_GREG_up - Wald_GREG_low;
            WALD_Bra_WIDTH_&i = Wald_Bra_up - Wald_Bra_low;
            IF Wilson_BASIC_low < 0 THEN WILSON_BASIC_INC_&i = 1;
            ELSE WILSON_BASIC_INC_&i = 0;
            IF Wald_GREG_low < 0 THEN WALD_GREG_INC_&i = 1;
            ELSE WALD_GREG_INC_&i = 0;
            IF Wald_Bra_low < 0 THEN WALD_Bra_INC_&i = 1;
            ELSE WALD_Bra_INC_&i = 0;
            %END;

RUN;
%MEND;
```

POPULATION9.sas

```
DATA POP1 (KEEP = L_pnr_UNIK ROBBERY -- INCOMECLASS);
            set DATA.NTU (keep=L_pnr_UNIK Insamlings_r C9 -- BB46_3 Kommun2 -- L_n
v_rldsdelnamnUP vlder CSFVI);

            if C9=2 then
            ROBBERY = 0;
            else if C9=1 then
            ROBBERY = 1;

            if C10=2 then
            ASSAULT = 0;
            else if C10=1 then
            ASSAULT = 1;

            if C10=2 then
            SEVERE_ASSAULT = 0;
            else if C10=1 AND (BB46_1=1 OR BB46_2=1 OR BB46_3=1) then
            SEVERE_ASSAULT = 1;
            else
            SEVERE_ASSAULT = 0;

            if C11=2 then
            SEXUAL = 0;
            else if C11=1 then
            SEXUAL = 1;

            if C12=2 then
            THREAT = 0;
            else if C12=1 then
            THREAT = 1;

            if C13=2 then
            FRAUD = 0;
            else if C13=1 then
            FRAUD = 1;

            if C14=2 then
```

```
            HARASS = 0;
            else if C14=1 then
            HARASS = 1;

            IF ROBBERY = 1 OR ASSAULT = 1 OR SEXUAL = 1 OR THREAT = 1 OR FRAUD = 1 OR HARASS
= 1 THEN
            ANY = 1;
            ELSE
            ANY = 0;

            if B210A_1=1 OR B210A_2=1 OR B210A_3=1 then
            ROBBERY_REPORT=1;
            else
            ROBBERY_REPORT=0;

            if B412A_1=1 OR B412A_2=1 OR B412A_3=1 then
            ASSAULT_REPORT=1;
            else
            ASSAULT_REPORT=0;

            if (B412A_1=1 AND BB46_1=1) OR (B412A_2=1 AND BB46_2=1) OR (B412A_2=1 AND
BB46_2=1) then
            SEVERE_ASSAULT_REPORT=1;
            else
            SEVERE_ASSAULT_REPORT=0;

            if B37A_1=1 OR B37A_2=1 OR B37A_3=1 then
            SEXUAL_REPORT=1;
            else
            SEXUAL_REPORT=0;

            if B510A_1=1 OR B510A_2=1 OR B510A_3=1 then
            THREAT_REPORT=1;
            else
            THREAT_REPORT=0;

            if B76A_1=1 OR B76A_1=1 OR B76A_1=1 then
            FRAUD_REPORT=1;
            else
            FRAUD_REPORT=0;

            if B66A=1 OR B66A=2 then
            HARASS_REPORT=1;
            else
            HARASS_REPORT=0;

            IF
            ROBBERY_REPORT = 1 OR
            ASSAULT_REPORT = 1 OR
            SEXUAL_REPORT = 1 OR
            THREAT_REPORT = 1 OR
            FRAUD_REPORT = 1 OR
            HARASS_REPORT = 1 THEN
            ANY_REPORT = 1;
            ELSE
            ANY_REPORT = 0;

            if Civil='G' OR Civil='RP' then
            CIVIL_STATUS=1;
            else
            CIVIL_STATUS=0;

    if  v_rldsdelnamnUP = 'Afrika' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'Asien' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'EU25 utom Norden' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'EU27 utom Norden' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'Europa utom EU25 och Norden' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'Europa utom EU27 och Norden' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'Nordamerika' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'Norden utom Sverige' then NON_NORDIC = 0;
    if  v_rldsdelnamnUP = 'Oceanien' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'Sovjetunionen' then NON_NORDIC = 1;
    if  v_rldsdelnamnUP = 'Sverige' then NON_NORDIC = 0;
    if  v_rldsdelnamnUP = 'Sydamerika' then NON_NORDIC = 1;

            if K_n='1' then
            SEX=0;
```

```
                else if K_n='2' then
                SEX=1;

                KOMMUN = INPUT(kommun2, 4.);
                LAN = L_n;

                AGE = vlder;

                if vlder<30 then
                AGECLASS=1;
                else if vlder>29 AND vlder<41 then
                AGECLASS=2;
                else if vlder>40 AND vlder<51 then
                AGECLASS=3;
                else if vlder>50 AND vlder<66 then
                AGECLASS=4;
                else if vlder>65 AND vlder<75 then
                AGECLASS=5;
                else if vlder>74 then
                AGECLASS=6;

                if vlder<30 then
                AGESTRATA = 1;
                else if vlder>29 AND vlder<75 then
                AGESTRATA=2;
                else if vlder>74 then
                AGESTRATA=3;

                INCOME = CSFVI;

                if CSFVI<150000 then
                INCOMECLASS=1;
                else if CSFVI>149999 AND CSFVI<300000 then
                INCOMECLASS=2;
                else if CSFVI>299999 then
                INCOMECLASS=3;

RUN;
PROC IMPORT
                DATAFILE =
                'H:\Måns\SAE\DATA\KOMMUNDATA.txt'
                OUT = DATA.KOMMUN
                DBMS=TAB
                REPLACE;
RUN;
PROC IMPORT
                DATAFILE =
                'H:\Måns\SAE\DATA\LANSDATA2.txt'
                OUT = DATA.LAN
                DBMS=TAB
                REPLACE;
RUN;
%MACRO ANMBROTT;
PROC SORT DATA=POP1;
                BY &NIVA;
RUN;
DATA TEMP1 (drop = NAMN -- _BEFOLKNING_ANTAL_AR);
                set DATA.&NIVA;
                ALLA_BROTT = MEAN(ALLA_2005,ALLA_2006,ALLA_2007,ALLA_2008);
                MISSHANDEL =
MEAN(MISSHANDEL_2005,MISSHANDEL_2006,MISSHANDEL_2007,MISSHANDEL_2008);
                HOT = MEAN(HOT_2005,HOT_2006,HOT_2007,HOT_2008);
                OFREDANDE = MEAN(OFREDANDE_2005,OFREDANDE_2006,OFREDANDE_2007,OFREDANDE_2008);
                SEXUALBROTT =
MEAN(SEXUALBROTT_2005,SEXUALBROTT_2006,SEXUALBROTT_2007,SEXUALBROTT_2008);
                PERSONRAN = MEAN(PERSONRAN_2005,PERSONRAN_2006,PERSONRAN_2007,PERSONRAN_2008);
                BEDRAGERI = MEAN(BEDRAGERI_2005,BEDRAGERI_2006,BEDRAGERI_2007,BEDRAGERI_2008);
                BEFOLKN_1NOV =
MEAN(BEFOLKN_1NOV_2005,BEFOLKN_1NOV_2006,BEFOLKN_1NOV_2007,BEFOLKN_1NOV_2008);
RUN;
PROC MEANS NOPRINT DATA=TEMP1;
                VAR ALLA_BROTT MISSHANDEL HOT OFREDANDE SEXUALBROTT PERSONRAN BEDRAGERI
BEFOLKN_1NOV;
                BY &NIVA_ART;
                OUTPUT OUT = TEMP2
                MEAN(ALLA_BROTT) = ALL_OFF_REPORTED_&NIVA_ART
                MEAN(PERSONRAN) = ROBBERY_REPORTED_&NIVA_ART
```

```
              MEAN(MISSHANDEL) = ASSAULT_REPORTED_&NIVA_ART
              MEAN(SEXUALBROTT) = SEXUAL_REPORTED_&NIVA_ART
              MEAN(HOT) = THREAT_REPORTED_&NIVA_ART
              MEAN(BEDRAGERI) = FRAUD_REPORTED_&NIVA_ART
              MEAN(OFREDANDE) = HARASS_REPORTED_&NIVA_ART
              MEAN(BEFOLKN_1NOV) = POPUL_1NOV_&NIVA_ART
              ;
RUN;
DATA TEMP3 (DROP = ALLA_BROTT -- POPUL_1NOV_&NIVA_ART);
              MERGE TEMP1 TEMP2;
              BY &NIVA_ART;
              ALL_OFF_REP_&NIVA_ART =
(100000/POPUL_1NOV_&NIVA_ART)*ALL_OFF_REPORTED_&NIVA_ART;
              ROBBERY_REP_&NIVA_ART =
(100000/POPUL_1NOV_&NIVA_ART)*ROBBERY_REPORTED_&NIVA_ART;
              ASSAULT_REP_&NIVA_ART =
(100000/POPUL_1NOV_&NIVA_ART)*ASSAULT_REPORTED_&NIVA_ART;
              SEVERE_ASSAULT_REP_&NIVA_ART = ASSAULT_REP_&NIVA_ART;
              SEXUAL_REP_&NIVA_ART = (100000/POPUL_1NOV_&NIVA_ART)*SEXUAL_REPORTED_&NIVA_ART;
              THREAT_REP_&NIVA_ART = (100000/POPUL_1NOV_&NIVA_ART)*THREAT_REPORTED_&NIVA_ART;
              FRAUD_REP_&NIVA_ART = (100000/POPUL_1NOV_&NIVA_ART)*FRAUD_REPORTED_&NIVA_ART;
              HARASS_REP_&NIVA_ART = (100000/POPUL_1NOV_&NIVA_ART)*HARASS_REPORTED_&NIVA_ART;
              ANY_REP_&NIVA_ART =
              (
              ROBBERY_REP_&NIVA_ART +
              ASSAULT_REP_&NIVA_ART +
              SEXUAL_REP_&NIVA_ART +
              THREAT_REP_&NIVA_ART +
              FRAUD_REP_&NIVA_ART +
              HARASS_REP_&NIVA_ART
              );

RUN;
DATA TEMP4;
              merge POP1 TEMP3;
              BY &NIVA;
RUN;
DATA POP1;
              set TEMP4;
RUN;
%MEND;
/* CREATING REPORTED CRIMES PER 100 000 INHABITANTS */
%let NIVA = KOMMUN;
%let NIVA_ART = MUNICIPALITY;
%anmbrott;
%let NIVA = LAN;
%let NIVA_ART = COUNTY;
%anmbrott;
DATA DATA.POPULATION (DROP = OVRIGT);
              set POP1;
              /* DEFINING STRATA IN SAMPLING DESIGN */
              if AGESTRATA=1 then
              STRATA_IND = COUNTY;
              else if AGESTRATA=2 then
              STRATA_IND = COUNTY+7;
              else if AGESTRATA=3 then
              STRATA_IND = COUNTY+14;

              /* DELETING PARTIAL NONRESPONSE */
              if MISSING(ROBBERY) then DELETE;
              if MISSING(ASSAULT) then DELETE;
              if MISSING(SEXUAL) then DELETE;
              if MISSING(THREAT) then DELETE;
              if MISSING(FRAUD) then DELETE;
              if MISSING(HARASS) then DELETE;
              if MISSING(ROBBERY_REPORT) then DELETE;
              if MISSING(ASSAULT_REPORT) then DELETE;
              if MISSING(SEXUAL_REPORT) then DELETE;
              if MISSING(THREAT_REPORT) then DELETE;
              if MISSING(FRAUD_REPORT) then DELETE;
              if MISSING(HARASS_REPORT) then DELETE;
              if MISSING(NON_NORDIC) then DELETE;

              COUNT_STR = STRATA_IND;
              if AGESTRATA=1 then
              MUNICIP_STR = MUNICIPALITY;
              else if AGESTRATA=2 then
```

```
            MUNICIP_STR = MUNICIPALITY+97;
            else if AGESTRATA=3 then
            MUNICIP_STR = MUNICIPALITY+194;
RUN;
PROC SORT DATA=DATA.POPULATION OUT=DATA.POP_SORT_STRATA; /* COPYING AND SORTING THE POPULATION
FILE TO FASTEN THE SIMULATION */
by STRATA_IND;
RUN;
PROC SORT DATA=DATA.POPULATION OUT=DATA.POP_SORT_NR; /* COPYING AND SORTING THE POPULATION
FILE TO FASTEN THE SIMULATION */
by L_pnr_UNIK;
RUN;
```

SAMPLE3.sas

```
PROC SURVEYSELECT DATA=DATA.POP_SORT_STRATA NOPRINT
            method=SRS
            n=(479 125 143 563 119 143 149 828 198 253 917 203 264 244 90 27 33 120 28 40
34)
    seed=&simu_nr
            out=SampleData;
            strata STRATA_IND;
RUN;
PROC SORT DATA=SampleData (KEEP = L_pnr_UNIK SelectionProb SamplingWeight) OUT=SampleData_min;
BY L_pnr_UNIK;
RUN;
DATA SIMULPOP_MODELFIT;
            merge DATA.POP_SORT_NR SampleData_min;
            BY L_pnr_UNIK;
            if MISSING(SamplingWeight) then
            SamplingWeight = 0;
            if SamplingWeight = 0 then
            SamplingIndicator = 0;
            else if SamplingWeight > 0 then
            SamplingIndicator = 1;
RUN;
```

STARTSETS5.sas

```
DATA RESULTAT.MODELS;
            length MODELNR $2. TYPE $6. RESPONSE $15. DEPENDENT $200.;
            INPUT MODELNR $ TYPE $ RESPONSE $ DEPENDENT $;
    datalines;
0 0 0 0
;
RUN;
DATA MODELS1;
            set RESULTAT.MODELS;
RUN;
DATA RESULTAT.CONVERGENCE;
RUN;
DATA SIMULINFO_BLANK;
            length SIMULERING 3. STARTTID 8. SLUTTID 8. SIMULERINGSTID 5.;
            INPUT SIMULERING STARTTID SLUTTID SIMULERINGSTID;
            format SIMULERINGSTID mmss5.0;
    datalines;
0 0 0 0
;
RUN;
DATA RESULTAT.SIMULINFO;
            SET SIMULINFO_BLANK;
RUN;
PROC SORT DATA=DATA.POPULATION OUT=POP_SORT_COUNTY;
by COUNTY;
RUN;
PROC MEANS DATA=POP_SORT_COUNTY NOPRINT; /* CALCULATING TRUE VALUES IN POPULATION */
            BY COUNTY;
            VAR ROBBERY -- ANY;
            OUTPUT
            OUT = TRUE_TOT_VALUES_COUNTY
            SUM(ROBBERY) = TRUE_TOT_ROBBERY
            SUM(ASSAULT) = TRUE_TOT_ASSAULT
            SUM(SEVERE_ASSAULT) = TRUE_TOT_SEVERE_ASSAULT
            SUM(SEXUAL) = TRUE_TOT_SEXUAL
            SUM(THREAT) = TRUE_TOT_THREAT
            SUM(FRAUD) = TRUE_TOT_FRAUD
            SUM(HARASS) = TRUE_TOT_HARASS
            SUM(ANY) = TRUE_TOT_ANY
```

```
            ;
RUN;
PROC SORT DATA=DATA.POPULATION OUT=POP_SORT_MUNICIPALITY;
by MUNICIPALITY;
RUN;
PROC MEANS DATA=POP_SORT_MUNICIPALITY NOPRINT; /* CALCULATING TRUE VALUES IN POPULATION */
            BY MUNICIPALITY;
            VAR ROBBERY -- ANY;
            OUTPUT
            OUT = TRUE_TOT_VALUES_MUNICIPALITY
            SUM(ROBBERY) = TRUE_TOT_ROBBERY
            SUM(ASSAULT) = TRUE_TOT_ASSAULT
            SUM(SEVERE_ASSAULT) = TRUE_TOT_SEVERE_ASSAULT
            SUM(SEXUAL) = TRUE_TOT_SEXUAL
            SUM(THREAT) = TRUE_TOT_THREAT
            SUM(FRAUD) = TRUE_TOT_FRAUD
            SUM(HARASS) = TRUE_TOT_HARASS
            SUM(ANY) = TRUE_TOT_ANY
            ;
RUN;
DATA TRUE_TOT_VALUES_MUNICIPALITY;
            SET TRUE_TOT_VALUES_MUNICIPALITY;
            LARGE_N = _FREQ_;
RUN;
DATA TRUE_TOT_VALUES_COUNTY;
            SET TRUE_TOT_VALUES_COUNTY;
            LARGE_N = _FREQ_;
RUN;
```

MACROANALYZE5.sas

```
%MACRO PROP;
PROC SORT DATA=RESULTAT.RESULTS_ALL;
            BY DOMAIN;
PROC MEANS DATA=RESULTAT.RESULTS_ALL NOPRINT;
            VAR SMALL_d_str_N;
            BY DOMAIN;
            OUTPUT OUT=TEMP N=FREQ MEAN(SMALL_d_str_N) = SMALL_d_str_N;
RUN;
DATA TEMP;
            SET TEMP;
            if SMALL_d_str_N <21 then DOMAINCLASS=1;
            else if SMALL_d_str_N <41 then DOMAINCLASS=2;
            else if SMALL_d_str_N <101 then DOMAINCLASS=3;
            else if SMALL_d_str_N <350 then DOMAINCLASS=4;
            else DOMAINCLASS=5;
            Proportion = FREQ/10;
RUN;
%MEND;

%MACRO VAR_HT_ZERO;
DATA TEMP (KEEP = DOMAIN SMALL_d_str_N DOMAINCLASS VAR_HT_&RESPONSE1._ZERO --
VAR_HT_&RESPONSE8._ZERO);
            SET RESULTAT.RESULTS_ALL;
            %DO j=1 %TO &RESPONSENR;
                        %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                        %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                        %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                        %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
                        %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
                        %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
                        %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
                        %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
                        IF VAR_HT_&RESPONSE > 0 THEN VAR_HT_&RESPONSE._ZERO = 0;
                        ELSE VAR_HT_&RESPONSE._ZERO = 1;
            %END;
RUN;

PROC SORT DATA=TEMP;
            BY DOMAIN;
RUN;

%DO j=1 %TO &RESPONSENR;
            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
```

```
            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
            PROC MEANS DATA=TEMP NOPRINT;
                        VAR SMALL_d_str_N VAR_HT_&RESPONSE._ZERO;
                        BY DOMAIN;
                        OUTPUT OUT=TEMP_&RESPONSE
                        MEAN(SMALL_d_str_N) = SMALL_d_str_N
                        MEAN(VAR_HT_&RESPONSE._ZERO) = ZERO
                        ;
            RUN;
            DATA TEMP_&RESPONSE;
                        SET TEMP_&RESPONSE;
                        if SMALL_d_str_N <21 then DOMAINCLASS=1;
                        else if SMALL_d_str_N <41 then DOMAINCLASS=2;
                        else if SMALL_d_str_N <101 then DOMAINCLASS=3;
                        else if SMALL_d_str_N <350 then DOMAINCLASS=4;
                        else DOMAINCLASS=5;
                        CRIME="                  ";
                        CRIME="&RESPONSE";
            RUN;
%END;

DATA TEMP2 (KEEP = HT_VAR_ZERO DOMAIN DOMAINCLASS CRIME);
            SET
            %DO j=1 %TO &RESPONSENR;
            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
            TEMP_&RESPONSE
            %END;
            ;
            HT_VAR_ZERO=ZERO*100;
RUN;
%MEND;

%MACRO GREG_NEGATIVE;
DATA TEMP (KEEP = DOMAIN SMALL_d_str_N DOMAINCLASS GREG_1_NEG -- GREG_&TOT_MODELS._NEG);
            SET RESULTAT.RESULTS_ALL;
            %DO i=1 %TO &TOT_MODELS;
            IF GREG_&i < 0 THEN GREG_&i._NEG = 1;
            ELSE GREG_&i._NEG = 0;
            %END;
RUN;

PROC SORT DATA=TEMP;
            BY DOMAIN;
RUN;

%DO i=1 %TO &TOT_MODELS;
            PROC MEANS DATA=TEMP NOPRINT;
                        VAR SMALL_d_str_N GREG_&i._NEG;
                        BY DOMAIN;
                        OUTPUT OUT=TEMP_&i
                        MEAN(SMALL_d_str_N) = SMALL_d_str_N
                        MEAN(GREG_&i._NEG) = GREG_NEG
                        ;
            RUN;
            DATA TEMP_&i;
                        SET TEMP_&i;
                        if SMALL_d_str_N <21 then DOMAINCLASS=1;
                        else if SMALL_d_str_N <41 then DOMAINCLASS=2;
                        else if SMALL_d_str_N <101 then DOMAINCLASS=3;
                        else if SMALL_d_str_N <350 then DOMAINCLASS=4;
                        else DOMAINCLASS=5;
                        MODELNR="  ";
                        MODELNR="&i";
            RUN;
%END;
```

```
DATA TEMP2 (KEEP = NEGATIVE_GREG DOMAIN DOMAINCLASS MODELNR);
              SET
              %DO i=1 %TO &TOT_MODELS;
              TEMP_&i
              %END;
              ;
              NEGATIVE_GREG=GREG_NEG*100;
RUN;

PROC SORT DATA=TEMP2;
              BY MODELNR;
RUN;

PROC SORT DATA=RESULTAT.MODELS;
              BY MODELNR;
RUN;

DATA TEMP3 (DROP = DEPENDENT);
              MERGE TEMP2 RESULTAT.MODELS;
              BY MODELNR;
RUN;
%MEND;

%MACRO AB_RMSE;
DATA TEMP (KEEP = DOMAIN SMALL_d_str_N DIFF_P_HT_&RESPONSE1 -- DIFFSQ_P_GREG_&TOT_MODELS);
              SET RESULTAT.RESULTS_ALL;
              %DO j=1 %TO &RESPONSENR;
              %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
              %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
              %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
              %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
              %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
              %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
              %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
              %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
              DIFF_P_HT_&RESPONSE = (HT_TOT_&RESPONSE/LARGE_N)-(TRUE_TOT_&RESPONSE/LARGE_N );
              DIFFSQ_P_HT_&RESPONSE = ((HT_TOT_&RESPONSE/LARGE_N)-(TRUE_TOT_&RESPONSE/LARGE_N
)) ** 2;
              %END;

              %DO i=1 %TO &TOT_MODELS;
                            %IF &i>=1 AND &i <= %eval(&DEPENDENTNR*2*1)
                                                                %THEN %LET RESPONSE =
&RESPONSE1;
                            %IF &i>=%eval(&DEPENDENTNR*2*1+1) AND &i<=%eval(&DEPENDENTNR*2*2)
              %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &i>=%eval(&DEPENDENTNR*2*2+1) AND &i<=%eval(&DEPENDENTNR*2*3)
              %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &i>=%eval(&DEPENDENTNR*2*3+1) AND &i<=%eval(&DEPENDENTNR*2*4)
              %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &i>=%eval(&DEPENDENTNR*2*4+1) AND &i<=%eval(&DEPENDENTNR*2*5)
              %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &i>=%eval(&DEPENDENTNR*2*5+1) AND &i<=%eval(&DEPENDENTNR*2*6)
              %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &i>=%eval(&DEPENDENTNR*2*6+1) AND &i<=%eval(&DEPENDENTNR*2*7)
              %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &i>=%eval(&DEPENDENTNR*2*7+1) AND &i<=%eval(&DEPENDENTNR*2*8)
              %THEN %LET RESPONSE = &RESPONSE8;

                            DIFF_P_GREG_&i = (GREG_&i/LARGE_N) - (TRUE_TOT_&RESPONSE/LARGE_N);
                            DIFFSQ_P_GREG_&i = ((GREG_&i/LARGE_N) -
(TRUE_TOT_&RESPONSE/LARGE_N))**2;

              %END;
RUN;

PROC SORT DATA=TEMP;
              BY DOMAIN;
RUN;

%DO j=1 %TO &RESPONSENR;
              %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
              %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
              %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
              %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
              %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
              %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
```

```
            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
            PROC MEANS DATA=TEMP NOPRINT;
                        VAR SMALL_d_str_N DIFF_P_HT_&RESPONSE DIFFSQ_P_HT_&RESPONSE;
                        BY DOMAIN;
                        OUTPUT OUT=TEMP_&RESPONSE
                        MEAN(SMALL_d_str_N) = SMALL_d_str_N
                        SUM(DIFF_P_HT_&RESPONSE) = SUM_DIFF_P_HT
                        SUM(DIFFSQ_P_HT_&RESPONSE) = SUM_DIFFSQ_P_HT
                        ;
            RUN;
            DATA TEMP_&RESPONSE;
                        SET TEMP_&RESPONSE;
                        if SMALL_d_str_N <21 then DOMAINCLASS=1;
                        else if SMALL_d_str_N <41 then DOMAINCLASS=2;
                        else if SMALL_d_str_N <101 then DOMAINCLASS=3;
                        else if SMALL_d_str_N <350 then DOMAINCLASS=4;
                        else DOMAINCLASS=5;
                        RESPONSE="                  ";
                        RESPONSE="&RESPONSE";
                        DEP=0;
                        TYPE="   ";
                        TYPE="HT";
            RUN;
%END;

DATA TEMP2 (KEEP = DOMAIN DOMAINCLASS RESPONSE DEP TYPE AB RMSE);
            SET
            %DO j=1 %TO &RESPONSENR;
            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
            TEMP_&RESPONSE
            %END;
            ;
            AB=abs(SUM_DIFF_P_HT/_FREQ_);
            RMSE=sqrt(SUM_DIFFSQ_P_HT/_FREQ_);
RUN;

%DO i=1 %TO &TOT_MODELS;
            PROC MEANS DATA=TEMP NOPRINT;
                        VAR SMALL_d_str_N DIFF_P_GREG_&i DIFFSQ_P_GREG_&i;
                        BY DOMAIN;
                        OUTPUT OUT=TEMP_&i
                        MEAN(SMALL_d_str_N) = SMALL_d_str_N
                        SUM(DIFF_P_GREG_&i) = SUM_DIFF_P_GREG
                        SUM(DIFFSQ_P_GREG_&i) = SUM_DIFFSQ_P_GREG
                        ;
            RUN;
            DATA TEMP_&i;
                        SET TEMP_&i;
                        if SMALL_d_str_N <21 then DOMAINCLASS=1;
                        else if SMALL_d_str_N <41 then DOMAINCLASS=2;
                        else if SMALL_d_str_N <101 then DOMAINCLASS=3;
                        else if SMALL_d_str_N <350 then DOMAINCLASS=4;
                        else DOMAINCLASS=5;
                        MODELNR="  ";
                        MODELNR="&i";
            RUN;
%END;

DATA TEMP3 (KEEP = DOMAIN DOMAINCLASS MODELNR AB RMSE);
            SET
            %DO i=1 %TO &TOT_MODELS;
            TEMP_&i
            %END;
            ;
            AB=abs(SUM_DIFF_P_GREG/_FREQ_);
            RMSE=sqrt(SUM_DIFFSQ_P_GREG/_FREQ_);
RUN;

PROC SORT DATA=TEMP3;
```

```
            BY MODELNR;
RUN;

PROC SORT DATA=RESULTAT.MODELS;
            BY MODELNR;
RUN;

DATA TEMP4 (DROP = DEPENDENT MODELNR);
            MERGE TEMP3 RESULTAT.MODELS;
            BY MODELNR;
RUN;

DATA TEMP5;
            SET TEMP4 TEMP2;
RUN;

DATA TEMP_DEP1 (KEEP = DOMAIN RESPONSE DOMAINCLASS RMSE1);
            SET TEMP5;
            IF DEP=1 AND TYPE='LIN';
            RMSE1=RMSE;
RUN;

DATA TEMP_DEP4 (KEEP = DOMAIN RESPONSE RMSE4);
            SET TEMP5;
            IF DEP=4 AND TYPE='LIN';
            RMSE4=RMSE;
RUN;

PROC SORT DATA=TEMP_DEP1;
            BY DOMAIN RESPONSE;
PROC SORT DATA=TEMP_DEP4;
            BY DOMAIN RESPONSE;
RUN;

DATA TEMP_DEP;
            MERGE TEMP_DEP1 TEMP_DEP4;
            BY DOMAIN RESPONSE;
RUN;

PROC SORT DATA=TEMP_DEP;
            BY DOMAINCLASS RESPONSE;
RUN;
PROC MEANS DATA=TEMP_DEP NOPRINT;
            VAR RMSE1 RMSE4;
            BY DOMAINCLASS RESPONSE;
            OUTPUT OUT=TEMP_DEP2
            MEAN(RMSE1) = RMSE1
            MEAN(RMSE4) = RMSE4
                        ;
RUN;

DATA TEMP_DEP3 (DROP = _TYPE_ _FREQ_ RMSE1 RMSE4);
            SET TEMP_DEP2;
            PROP_DIFF=RMSE4/RMSE1;
RUN;

%MEND;
%MACRO VAR_P_U;
DATA TEMP (KEEP = DOMAIN SMALL_d_str_N DOMAINCLASS DIFF_VAR_1 -- DIFF_VAR_&TOT_MODELS);
            SET RESULTAT.RESULTS_ALL;
            %DO i=1 %TO &TOT_MODELS;
            DIFF_VAR_&i = VAR_U_GREG_&i / VAR_P_GREG_&i;
            %END;
RUN;

PROC SORT DATA=TEMP;
            BY DOMAIN;
RUN;

%DO i=1 %TO &TOT_MODELS;
            PROC MEANS DATA=TEMP NOPRINT;
                        VAR SMALL_d_str_N DIFF_VAR_&i;
                        BY DOMAIN;
                        OUTPUT OUT=TEMP_&i
                        MEAN(SMALL_d_str_N) = SMALL_d_str_N
                        MEAN(DIFF_VAR_&i) = DIFF_VAR
                        ;
```

```
            RUN;
            DATA TEMP_&i;
                        SET TEMP_&i;
                        if SMALL_d_str_N <21 then DOMAINCLASS=1;
                        else if SMALL_d_str_N <41 then DOMAINCLASS=2;
                        else if SMALL_d_str_N <101 then DOMAINCLASS=3;
                        else if SMALL_d_str_N <350 then DOMAINCLASS=4;
                        else DOMAINCLASS=5;
                        MODELNR="  ";
                        MODELNR="&i";
            RUN;
%END;

DATA TEMP2 (KEEP = NEGATIVE_GREG DOMAIN DOMAINCLASS MODELNR DIFF_VAR);
            SET
            %DO i=1 %TO &TOT_MODELS;
            TEMP_&i
            %END;
            ;
RUN;

PROC SORT DATA=TEMP2;
            BY MODELNR;
RUN;

PROC SORT DATA=RESULTAT.MODELS;
            BY MODELNR;
RUN;

DATA TEMP3 (DROP = DEPENDENT);
            MERGE TEMP2 RESULTAT.MODELS;
            BY MODELNR;
RUN;
%MEND;
%MACRO INTERVAL;
DATA TEMP (KEEP = TRUE_TOT_ROBBERY -- TRUE_TOT_ANY DOMAIN SMALL_d_str_N LARGE_d_str_N
WALD_HT_COV_&RESPONSE1 -- WILSON_GREG_LEN_&TOT_MODELS);
            SET RESULTAT.RESULTS_ALL;
            %DO j=1 %TO &RESPONSENR;
            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
            IF
            TRUE_TOT_&RESPONSE < HT_TOT_&RESPONSE + &ALPHA * sqrt(VAR_HT_&RESPONSE)
            AND
            TRUE_TOT_&RESPONSE > HT_TOT_&RESPONSE - &ALPHA * sqrt(VAR_HT_&RESPONSE)
            THEN WALD_HT_COV_&RESPONSE = 1;
            ELSE WALD_HT_COV_&RESPONSE = 0;

            WALD_HT_LEN_&RESPONSE =
            (
            (HT_TOT_&RESPONSE/LARGE_N + &ALPHA * sqrt(VAR_HT_&RESPONSE/(LARGE_N**2)))
            -
            (HT_TOT_&RESPONSE/LARGE_N - &ALPHA * sqrt(VAR_HT_&RESPONSE/(LARGE_N**2)))
            )*100;

            IF
            HT_TOT_&RESPONSE + &ALPHA * sqrt(VAR_HT_&RESPONSE) > LARGE_d_str_N
            OR
            HT_TOT_&RESPONSE - &ALPHA * sqrt(VAR_HT_&RESPONSE) < 0
            THEN WALD_HT_INC_&RESPONSE = 1;
            ELSE WALD_HT_INC_&RESPONSE = 0;

            IF
            (TRUE_TOT_&RESPONSE/LARGE_N) <
            (
            (HT_TOT_&RESPONSE/LARGE_N) + (&ALPHA**2)/(2*SMALL_d_str_N)
            +
            &ALPHA * sqrt(
            ((HT_TOT_&RESPONSE/LARGE_N)*(1-
(HT_TOT_&RESPONSE/LARGE_N))/SMALL_d_str_N)+(&ALPHA**2)/(4*(SMALL_d_str_N**2))
            )
```

```
              )
              /
              (1+(&ALPHA**2)/SMALL_d_str_N)
              AND
              (TRUE_TOT_&RESPONSE/LARGE_N) >
              (
              (HT_TOT_&RESPONSE/LARGE_N) + (&ALPHA**2)/(2*SMALL_d_str_N)
              -
              &ALPHA * sqrt(
              ((HT_TOT_&RESPONSE/LARGE_N)*(1-
(HT_TOT_&RESPONSE/LARGE_N))/SMALL_d_str_N)+(&ALPHA**2)/(4*(SMALL_d_str_N**2))
              )
              )
              /
              (1+(&ALPHA**2)/SMALL_d_str_N)
              THEN WILSON_HT_COV_&RESPONSE = 1;
              ELSE WILSON_HT_COV_&RESPONSE = 0;

              WILSON_HT_LEN_&RESPONSE =
              (
              (
              (HT_TOT_&RESPONSE/LARGE_N) + (&ALPHA**2)/(2*SMALL_d_str_N)
              +
              &ALPHA * sqrt(
              ((HT_TOT_&RESPONSE/LARGE_N)*(1-
(HT_TOT_&RESPONSE/LARGE_N))/SMALL_d_str_N)+(&ALPHA**2)/(4*(SMALL_d_str_N**2))
              )
              )
              /
              (1+(&ALPHA**2)/SMALL_d_str_N)
              -
              (
              (HT_TOT_&RESPONSE/LARGE_N) + (&ALPHA**2)/(2*SMALL_d_str_N)
              -
              &ALPHA * sqrt(
              ((HT_TOT_&RESPONSE/LARGE_N)*(1-
(HT_TOT_&RESPONSE/LARGE_N))/SMALL_d_str_N)+(&ALPHA**2)/(4*(SMALL_d_str_N**2))
              )
              )
              /
              (1+(&ALPHA**2)/SMALL_d_str_N)
              )*100
              ;

              P_TRUE_&RESPONSE=(TRUE_TOT_&RESPONSE/LARGE_N);
              %END;

              %DO i=1 %TO &TOT_MODELS;
                            %IF &i>=1 AND &i <= %eval(&DEPENDENTNR*2*1)
                                                                  %THEN %LET RESPONSE =
&RESPONSE1;
                            %IF &i>=%eval(&DEPENDENTNR*2*1+1) AND &i<=%eval(&DEPENDENTNR*2*2)
              %THEN %LET RESPONSE = &RESPONSE2;
                            %IF &i>=%eval(&DEPENDENTNR*2*2+1) AND &i<=%eval(&DEPENDENTNR*2*3)
              %THEN %LET RESPONSE = &RESPONSE3;
                            %IF &i>=%eval(&DEPENDENTNR*2*3+1) AND &i<=%eval(&DEPENDENTNR*2*4)
              %THEN %LET RESPONSE = &RESPONSE4;
                            %IF &i>=%eval(&DEPENDENTNR*2*4+1) AND &i<=%eval(&DEPENDENTNR*2*5)
              %THEN %LET RESPONSE = &RESPONSE5;
                            %IF &i>=%eval(&DEPENDENTNR*2*5+1) AND &i<=%eval(&DEPENDENTNR*2*6)
              %THEN %LET RESPONSE = &RESPONSE6;
                            %IF &i>=%eval(&DEPENDENTNR*2*6+1) AND &i<=%eval(&DEPENDENTNR*2*7)
              %THEN %LET RESPONSE = &RESPONSE7;
                            %IF &i>=%eval(&DEPENDENTNR*2*7+1) AND &i<=%eval(&DEPENDENTNR*2*8)
              %THEN %LET RESPONSE = &RESPONSE8;

              %IF &NEG_CORRECT=1 %THEN
              %DO;
                            IF GREG_&i<0 THEN GREG_&i=0;
              %END;

              IF
              TRUE_TOT_&RESPONSE < GREG_&i + &ALPHA * sqrt(VAR_&VAR_TYPE._GREG_&i)
              AND
              TRUE_TOT_&RESPONSE > GREG_&i - &ALPHA * sqrt(VAR_&VAR_TYPE._GREG_&i)
              THEN WALD_GREG_COV_&i = 1;
              ELSE WALD_GREG_COV_&i = 0;
```

```
                IF
                GREG_&i + &ALPHA * sqrt(VAR_&VAR_TYPE._GREG_&i) > LARGE_d_str_N
                OR
                GREG_&i - &ALPHA * sqrt(VAR_&VAR_TYPE._GREG_&i) < 0
                THEN WALD_GREG_INC_&i = 1;
                ELSE WALD_GREG_INC_&i = 0;

                WALD_GREG_LEN_&i =
                (
                ((GREG_&i/LARGE_N) + &ALPHA * sqrt(VAR_&VAR_TYPE._GREG_&i/(LARGE_N**2)))
                -
                ((GREG_&i/LARGE_N) - &ALPHA * sqrt(VAR_&VAR_TYPE._GREG_&i/(LARGE_N**2)))
                )*100
                ;

                IF
                (TRUE_TOT_&RESPONSE/LARGE_N) <
                (
                (GREG_&i/LARGE_N) + (&ALPHA**2)/(2*SMALL_d_str_N)
                +
                &ALPHA * sqrt(
                (VAR_&VAR_TYPE._GREG_&i / LARGE_N**2) +(&ALPHA**2)/(4*(SMALL_d_str_N**2))
                )
                )
                /
                (1+(&ALPHA**2)/SMALL_d_str_N)
                AND
                (TRUE_TOT_&RESPONSE/LARGE_N) >
                (
                (GREG_&i/LARGE_N) + (&ALPHA**2)/(2*SMALL_d_str_N)
                -
                &ALPHA * sqrt(
                (VAR_&VAR_TYPE._GREG_&i / LARGE_N**2) +(&ALPHA**2)/(4*(SMALL_d_str_N**2))
                )
                )
                /
                (1+(&ALPHA**2)/SMALL_d_str_N)
                THEN WILSON_GREG_COV_&i = 1;
                ELSE WILSON_GREG_COV_&i = 0;

                WILSON_GREG_LEN_&i =
                (
                (
                (GREG_&i/LARGE_N) + (&ALPHA**2)/(2*SMALL_d_str_N)
                +
                &ALPHA * sqrt(
                (VAR_&VAR_TYPE._GREG_&i / LARGE_N**2) +(&ALPHA**2)/(4*(SMALL_d_str_N**2))
                )
                )
                /
                (1+(&ALPHA**2)/SMALL_d_str_N)
                -
                (
                (GREG_&i/LARGE_N) + (&ALPHA**2)/(2*SMALL_d_str_N)
                -
                &ALPHA * sqrt(
                (VAR_&VAR_TYPE._GREG_&i / LARGE_N**2) +(&ALPHA**2)/(4*(SMALL_d_str_N**2))
                )
                )
                /
                (1+(&ALPHA**2)/SMALL_d_str_N)
                )*100
                ;

                %END;
RUN;

PROC SORT DATA=TEMP;
                BY DOMAIN;
RUN;

%DO j=1 %TO &RESPONSENR;
                %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
                %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
                %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
                %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
```

```
            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
            PROC MEANS DATA=TEMP NOPRINT;
                        VAR SMALL_d_str_N WALD_HT_COV_&RESPONSE WALD_HT_INC_&RESPONSE
                                    WILSON_HT_COV_&RESPONSE TRUE_TOT_&RESPONSE
P_TRUE_&RESPONSE;
                        BY DOMAIN;
                        OUTPUT OUT=TEMP_&RESPONSE
                        MEAN(SMALL_d_str_N) = SMALL_d_str_N
                        MEAN(WALD_HT_COV_&RESPONSE) = WALD_COV
                        MEAN(WALD_HT_INC_&RESPONSE) = WALD_INC
                        MEAN(WALD_HT_LEN_&RESPONSE) = WALD_LEN
                        MEAN(WILSON_HT_COV_&RESPONSE) = WILSON_COV
                        MEAN(WILSON_HT_LEN_&RESPONSE) = WILSON_LEN
                        MEAN(TRUE_TOT_&RESPONSE) = TRUE_TOT_VALUE
                        MEAN(P_TRUE_&RESPONSE) = TRUE_P_VALUE
                        ;
            RUN;
            DATA TEMP_&RESPONSE;
                        SET TEMP_&RESPONSE;
                        if SMALL_d_str_N <21 then DOMAINCLASS=1;
                        else if SMALL_d_str_N <41 then DOMAINCLASS=2;
                        else if SMALL_d_str_N <101 then DOMAINCLASS=3;
                        else if SMALL_d_str_N <350 then DOMAINCLASS=4;
                        else DOMAINCLASS=5;
                        RESPONSE="                ";
                        RESPONSE="&RESPONSE";
                        DEP=0;
                        TYPE="   ";
                        TYPE="HT";
            RUN;
%END;

DATA TEMP2 (KEEP = DOMAIN DOMAINCLASS RESPONSE DEP TYPE WALD_COV WALD_LEN
                                    WALD_INC WILSON_COV WILSON_LEN TRUE_TOT_VALUE
TRUE_P_VALUE);
            SET
            %DO j=1 %TO &RESPONSENR;
            %IF &j=1 %THEN %LET RESPONSE = &RESPONSE1;
            %IF &j=2 %THEN %LET RESPONSE = &RESPONSE2;
            %IF &j=3 %THEN %LET RESPONSE = &RESPONSE3;
            %IF &j=4 %THEN %LET RESPONSE = &RESPONSE4;
            %IF &j=5 %THEN %LET RESPONSE = &RESPONSE5;
            %IF &j=6 %THEN %LET RESPONSE = &RESPONSE6;
            %IF &j=7 %THEN %LET RESPONSE = &RESPONSE7;
            %IF &j=8 %THEN %LET RESPONSE = &RESPONSE8;
            TEMP_&RESPONSE
            %END;
            ;
RUN;

%DO i=1 %TO &TOT_MODELS;
            %IF &i>=1 AND &i <= %eval(&DEPENDENTNR*2*1)
                                                %THEN %LET RESPONSE = &RESPONSE1;
            %IF &i>=%eval(&DEPENDENTNR*2*1+1) AND &i<=%eval(&DEPENDENTNR*2*2)   %THEN %LET
RESPONSE = &RESPONSE2;
            %IF &i>=%eval(&DEPENDENTNR*2*2+1) AND &i<=%eval(&DEPENDENTNR*2*3)   %THEN %LET
RESPONSE = &RESPONSE3;
            %IF &i>=%eval(&DEPENDENTNR*2*3+1) AND &i<=%eval(&DEPENDENTNR*2*4)   %THEN %LET
RESPONSE = &RESPONSE4;
            %IF &i>=%eval(&DEPENDENTNR*2*4+1) AND &i<=%eval(&DEPENDENTNR*2*5)   %THEN %LET
RESPONSE = &RESPONSE5;
            %IF &i>=%eval(&DEPENDENTNR*2*5+1) AND &i<=%eval(&DEPENDENTNR*2*6)   %THEN %LET
RESPONSE = &RESPONSE6;
            %IF &i>=%eval(&DEPENDENTNR*2*6+1) AND &i<=%eval(&DEPENDENTNR*2*7)   %THEN %LET
RESPONSE = &RESPONSE7;
            %IF &i>=%eval(&DEPENDENTNR*2*7+1) AND &i<=%eval(&DEPENDENTNR*2*8)   %THEN %LET
RESPONSE = &RESPONSE8;
            PROC MEANS DATA=TEMP NOPRINT;
                        VAR SMALL_d_str_N WALD_GREG_COV_&i WALD_GREG_INC_&i
WILSON_GREG_COV_&i
                                    WALD_GREG_LEN_&i WILSON_GREG_LEN_&i
TRUE_TOT_&RESPONSE P_TRUE_&RESPONSE;
                        BY DOMAIN;
                        OUTPUT OUT=TEMP_&i
```

```
                        MEAN(SMALL_d_str_N) = SMALL_d_str_N
                        MEAN(WALD_GREG_COV_&i) = WALD_COV
                        MEAN(WALD_GREG_INC_&i) = WALD_INC
                        MEAN(WALD_GREG_LEN_&i) = WALD_LEN
                        MEAN(WILSON_GREG_COV_&i) = WILSON_COV
                        MEAN(WILSON_GREG_LEN_&i) = WILSON_LEN
                        MEAN(TRUE_TOT_&RESPONSE) = TRUE_TOT_VALUE
                        MEAN(P_TRUE_&RESPONSE) = TRUE_P_VALUE
                        ;
            RUN;
            DATA TEMP_&i;
                        SET TEMP_&i;
                        if SMALL_d_str_N <21 then DOMAINCLASS=1;
                        else if SMALL_d_str_N <41 then DOMAINCLASS=2;
                        else if SMALL_d_str_N <101 then DOMAINCLASS=3;
                        else if SMALL_d_str_N <350 then DOMAINCLASS=4;
                        else DOMAINCLASS=5;
                        MODELNR="  ";
                        MODELNR="&i";
            RUN;
%END;

DATA TEMP3 (KEEP = DOMAIN DOMAINCLASS MODELNR WALD_COV WALD_LEN WALD_INC
                                    WILSON_COV WILSON_LEN TRUE_TOT_VALUE TRUE_P_VALUE);
            SET
            %DO i=1 %TO &TOT_MODELS;
            TEMP_&i
            %END;
            ;
RUN;

PROC SORT DATA=TEMP3;
            BY MODELNR;
RUN;

PROC SORT DATA=RESULTAT.MODELS;
            BY MODELNR;
RUN;

DATA TEMP4 (DROP = DEPENDENT MODELNR);
            MERGE TEMP3 RESULTAT.MODELS;
            BY MODELNR;
RUN;

DATA TEMP5 ;
            SET TEMP4 TEMP2;
            IF TRUE_P_VALUE=0 THEN ZEROVALUE=1;
            ELSE ZEROVALUE=0;
            IF TRUE_P_VALUE=0 THEN TOTVALUE=0;
            IF TRUE_P_VALUE<3 AND TRUE_P_VALUE>0 THEN TOTVALUE=1;
            IF TRUE_P_VALUE>2 THEN TOTVALUE=2;
RUN;

DATA TEMP6 ;
            SET TEMP5;
            IF DOMAINCLASS>3 THEN DELETE;
            IF RESPONSE='ROBBERY' OR RESPONSE='SEXUAL';
RUN;
DATA TEMP7 ;
            SET TEMP5;
            IF DEP=2 OR DEP=3 THEN DELETE;
            IF TYPE='LOG' THEN DELETE;
RUN;

%MEND;
```

ANALYZE9.sas

```
* ========== LIBNAMES ========== */
libname DATA 'H:\Måns\SAE\DATA';
libname RESULT1 'H:\Måns\SAE\RESULTAT_1_50';
libname RESULT2 'H:\Måns\SAE\RESULTAT_51_100';
libname RESULT3 'H:\Måns\SAE\RESULTAT_101_150';
libname RESULT4 'H:\Måns\SAE\RESULTAT_151_200';
libname RESULT5 'H:\Måns\SAE\RESULTAT_201_300';
libname RESULT6 'H:\Måns\SAE\RESULTAT_301_600';
libname RESULT7 'H:\Måns\SAE\RESULTAT_601_650';
```

```
libname RESULT8 'H:\Måns\SAE\RESULTAT_651_750';
libname RESULT9 'H:\Måns\SAE\RESULTAT_751_850';
libname RESULT10 'H:\Måns\SAE\RESULTAT_851_1000';
libname RESULTAT 'H:\Måns\SAE\RESULTAT_ALL';

/* READING MACROS */
%include 'H:\Måns\SAE\KOD\MACROSANALYZE5.sas';
/* PUTTING ALL SIMULATION RESULTS TOGETHER AS ONE FILE */
DATA RESULTAT.ESTIMATES_COUNTY_ALL;
              SET
              RESULT1.Estimates_county_1_50
              RESULT2.Estimates_county_51_100
              RESULT3.Estimates_county_101_150
              RESULT4.Estimates_county_151_200
              RESULT5.Estimates_county_201_300
              RESULT6.Estimates_county_301_600
              RESULT7.Estimates_county_601_650
              RESULT8.Estimates_county_651_750
              RESULT9.Estimates_county_751_850
              RESULT10.Estimates_county_851_1000
              ;
RUN;
DATA RESULTAT.ESTIMATES_MUNICIPALITY_ALL;
              SET
              RESULT1.Estimates_municip_1_50
              RESULT2.Estimates_municip_51_100
              RESULT3.Estimates_municip_101_150
              RESULT4.Estimates_municip_151_200
              RESULT5.Estimates_municip_201_300
              RESULT6.Estimates_municip_301_600
              RESULT7.Estimates_municip_601_650
              RESULT8.Estimates_municip_651_750
              RESULT9.Estimates_municip_751_850
              RESULT10.Estimates_municip_851_1000
              ;
RUN;
DATA RESULTAT.Convergence_ALL;
              SET
              RESULT1.convergence
              RESULT2.convergence
              RESULT3.convergence
              RESULT4.convergence
              RESULT5.convergence
              RESULT6.convergence
              RESULT7.convergence
              RESULT8.convergence
              RESULT9.convergence
              RESULT10.convergence
              ;
              IF Status=. THEN DELETE;
RUN;
DATA RESULTAT.SIMULINFO_ALL;
              SET
              RESULT1.SIMULINFO
              RESULT2.SIMULINFO
              RESULT3.SIMULINFO
              RESULT4.SIMULINFO
              RESULT5.SIMULINFO
              RESULT6.SIMULINFO
              RESULT7.SIMULINFO
              RESULT8.SIMULINFO
              RESULT9.SIMULINFO
              RESULT10.SIMULINFO
              ;
              IF SIMULERING=0 THEN DELETE;
RUN;
DATA RESULTAT.MODELS;
              SET
              RESULT1.MODELS
              ;
              IF DEPENDENT =
              'COUNTY SEX AGECLASS NON_NORDIC CIVIL_STATUS INCOMECLASS URBAN'
              THEN DEP = 1;
              ELSE IF DEPENDENT =
              'MUNICIPALITY SEX AGECLASS NON_NORDIC CIVIL_STATUS INCOMECLASS URBAN'
              THEN DEP = 3;
              ELSE IF LENGTH(DEPENDENT) > 80 THEN DEP = 2;
```

```
            ELSE IF LENGTH(DEPENDENT) < 80 THEN DEP = 4;
RUN;
DATA RESULTAT.RESULTS_ALL (DROP = _TYPE_ _FREQ_);
            set RESULTAT.ESTIMATES_COUNTY_ALL RESULTAT.ESTIMATES_MUNICIPALITY_ALL;
            IF COUNTY =. THEN DOMAIN=MUNICIPALITY;
            ELSE DOMAIN = COUNTY * 100;
            if SMALL_d_str_N <21 then DOMAINCLASS=1;
            else if SMALL_d_str_N <41 then DOMAINCLASS=2;
            else if SMALL_d_str_N <101 then DOMAINCLASS=3;
            else if SMALL_d_str_N <350 then DOMAINCLASS=4;
            else DOMAINCLASS=5;
            T=quantile('T',.975,SMALL_d_str_N);
            Z=quantile('NORMAL',.975);
RUN;
PROC SORT DATA=RESULTAT.RESULTS_ALL;
            BY DOMAIN;
RUN;

PROC FORMAT;
            VALUE DOMAINCLASS          1 = '    -  20'
                                                                        2 = ' 21 -
40'
                                                                        3 = ' 41 -
100'
                                                                        4 = '101 -
349'
                                                                        5 = '350 -
'
                                                                        ;
            VALUE $TYPE                'LOG' = 'Logistic GREG'
                                                                    'LIN' = 'Linear GREG'
                                                                    'HT' = 'HT estimator'
                                                                    ;
            VALUE $RESPONSE            'ROBBERY' = 'Robbery'
                                                                    'ASSAULT' = 'Assault'
                                                                    'SEVERE_ASSAULT' = 'Severe
assault'
                                                                    'SEXUAL' = 'Sexual
offences'
                                                                    'THREAT' = 'Threat'
                                                                    'FRAUD' = 'Fraud'
                                                                    'HARASS' = 'Harassment'
                                                                    'ANY' = 'Any crime against
person'
                                                                    ;
            VALUE         DEP          0 = 'None'
                                                                    1 = 'Setup 1'
                                                                    2 = 'Setup 2'
                                                                    3 = 'Setup 3'
                                                                    4 = 'Setup 4'
                                                                    ;
            VALUE ZEROVALUE            1 = '  0'
                                                                        0 = '> 0'
                                                                        ;
RUN;
/* ANALYZE THE PROPORTION OF DOMAINS THAT CAN BE ESTIMATED */
/* CREATES A DATASET PROP*/
%PROP
PROC TABULATE DATA=TEMP;
            CLASS DOMAINCLASS;
            FORMAT DOMAINCLASS DOMAINCLASS.;
            VAR Proportion;
            TABLE DOMAINCLASS=''*Proportion='',(N='Number of domains'*FORMAT=2.0
MEAN='Percent (%)'*FORMAT=5.1)
            /BOX='Mean sample size in domain' ROW=FLOAT;
            TITLE 'Percentage of domains where all strata have a sample size greater than
1';
RUN;
/* ANALYZE ZERO VAR HT-ESTIMATES */
/* TABULATE RESULTS*/
%VAR_HT_ZERO
PROC TABULATE DATA=TEMP2;
            CLASS CRIME DOMAINCLASS;
            FORMAT DOMAINCLASS DOMAINCLASS. CRIME $RESPONSE.;
            VAR HT_VAR_ZERO;
            TABLE CRIME='', DOMAINCLASS='Mean sample size in
domain'*HT_VAR_ZERO=''*MEAN=''*FORMAT=5.1
```

```
             /BOX='Type of experienced offence' ROW=FLOAT;
             TITLE 'Percentage of Horwitz-Thompson variance estimate of zero';
RUN;
PROC TABULATE DATA=RESULTAT.Convergence_all;
             CLASS Status Model;
             TABLE Model, Status
             /BOX='Type of experienced offence' ROW=FLOAT;
             TITLE 'Test';
RUN;
/* ANALYZE NEGATIVE GREG-ESTIMATES */
%GREG_NEGATIVE

PROC TABULATE DATA=TEMP3;
             CLASS TYPE RESPONSE DEP DOMAINCLASS;
             FORMAT
                          DOMAINCLASS DOMAINCLASS.
                          RESPONSE $RESPONSE.
                          DEP DEP.
                          TYPE $TYPE.
                          ;
             VAR NEGATIVE_GREG;
             TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
             MEAN=''*NEGATIVE_GREG=''*RESPONSE=''*FORMAT=5.1
             /ROW=FLOAT;
             TITLE 'Percentage of Negative GREG estimates';
RUN;

/* ANALYZE AVERAGE BIAS AND RMSE */
%AB_RMSE
PROC TABULATE DATA=TEMP5;
             CLASS TYPE RESPONSE DEP DOMAINCLASS;
             FORMAT
                          DOMAINCLASS DOMAINCLASS.
                          RESPONSE $RESPONSE.
                          DEP DEP.
                          TYPE $TYPE.
                          ;
             VAR AB;
             TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
             MEAN=''*AB=''*RESPONSE=''*FORMAT=5.3
             /ROW=FLOAT;
             TITLE 'Average Bias (AB)';
RUN;
PROC TABULATE DATA=TEMP5;
             CLASS TYPE RESPONSE DEP DOMAINCLASS;
             FORMAT
                          DOMAINCLASS DOMAINCLASS.
                          RESPONSE $RESPONSE.
                          DEP DEP.
                          TYPE $TYPE.
                          ;
             VAR RMSE;
             TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
             MEAN=''*RMSE=''*RESPONSE=''*FORMAT=5.3
             /ROW=FLOAT;
             TITLE 'Root mean squared error (RMSE)';
RUN;
PROC TABULATE DATA=TEMP_DEP3;
             CLASS RESPONSE DOMAINCLASS;
             FORMAT
                          DOMAINCLASS DOMAINCLASS.
                          RESPONSE $RESPONSE.
                          ;
             VAR PROP_DIFF;
             TABLE DOMAINCLASS='Mean sample size in domain' ALL,
             MEAN=''*PROP_DIFF=''*RESPONSE=''*FORMAT=5.3
             /ROW=FLOAT;
             TITLE 'Proportional difference in RMSE between GREG (linear) Setup 1 and GREG
(linear) Setup 4';
RUN;

/* ANALYZE DIFFERENCE BETWEEN VAR_P AND VAR_U */
%VAR_P_U
PROC TABULATE DATA=TEMP3;
```

```
              CLASS TYPE RESPONSE DEP DOMAINCLASS;
              FORMAT
                           DOMAINCLASS DOMAINCLASS.
                           RESPONSE $RESPONSE.
                           DEP DEP.
                           TYPE $TYPE.
                           ;
              VAR DIFF_VAR;
              TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
              MEAN=''*DIFF_VAR=''*RESPONSE=''*FORMAT=5.1
              /ROW=FLOAT;
              TITLE 'Proportional difference between planned domain variance and unplanned
domain variance';
RUN;


/* ANALYZE INTERVAL COVERAGE AND INCORRECT INTERVALS */
/* WHICH VARIANCE TO USE IN CALCULATIONS P=Planned U=Unplanned:*/
%include 'H:\Måns\SAE\KOD\MACROSANALYZE5.sas';
%LET VAR_TYPE=P;
/* Correct negative GREG estimates to zero 1=YES, 0=No:*/
%LET NEG_CORRECT=1;
/* DISTRIBUTION IN INTERVAL ESTIMATES: T o (T) or Z (Normal)*/
%LET ALPHA=T;

%INTERVAL
PROC TABULATE DATA=TEMP5;
              CLASS TYPE RESPONSE DEP DOMAINCLASS;
              FORMAT
                           DOMAINCLASS DOMAINCLASS.
                           RESPONSE $RESPONSE.
                           DEP DEP.
                           TYPE $TYPE.
                           ;
              VAR WALD_COV;
              TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
              MEAN=''*WALD_COV=''*RESPONSE=''*FORMAT=5.3
              /ROW=FLOAT;
              TITLE 'Wald interval coverage percentage (95%, z=1.96)';
RUN;
PROC TABULATE DATA=TEMP5;
              CLASS TYPE RESPONSE DEP DOMAINCLASS;
              FORMAT
                           DOMAINCLASS DOMAINCLASS.
                           RESPONSE $RESPONSE.
                           DEP DEP.
                           TYPE $TYPE.
                           ;
              VAR WALD_INC;
              TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
              MEAN=''*WALD_INC=''*RESPONSE=''*FORMAT=5.3
              /ROW=FLOAT;
              TITLE 'Wald incorrect interval percentage (incudes negative values)';
RUN;
PROC TABULATE DATA=TEMP5;
              CLASS TYPE RESPONSE DEP DOMAINCLASS;
              FORMAT
                           DOMAINCLASS DOMAINCLASS.
                           RESPONSE $RESPONSE.
                           DEP DEP.
                           TYPE $TYPE.
                           ;
              VAR WILSON_COV;
              TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
              MEAN=''*WILSON_COV=''*RESPONSE=''*FORMAT=5.3
              /ROW=FLOAT;
              TITLE 'Wilson coverage percentage (95%)';
RUN;
PROC TABULATE DATA=TEMP6;
              CLASS TYPE RESPONSE DEP DOMAINCLASS TOT_VALUE;
              FORMAT
                           DOMAINCLASS DOMAINCLASS.
                           RESPONSE $RESPONSE.
```

```
                        DEP DEP.
                        TYPE $TYPE.
                        ;
            VAR WILSON_COV;
            TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
            MEAN=''*WILSON_COV=''*RESPONSE=''*TOT_VALUE='True value in
population'*FORMAT=5.3
            /ROW=FLOAT MISSTEXT='*';
            TITLE 'Wilson coverage for more rare events (95%)';
RUN;
PROC TABULATE DATA=TEMP7;
            CLASS TYPE RESPONSE DEP DOMAINCLASS ;
            FORMAT
                        DOMAINCLASS DOMAINCLASS.
                        RESPONSE $RESPONSE.
                        DEP DEP.
                        TYPE $TYPE.
                        ZEROVALUE ZEROVALUE.
                        ;
            VAR WILSON_LEN WALD_LEN;
            TABLE DOMAINCLASS='Mean sample size in domain'*DEP='Auxiliary
variables'*TYPE='Estimator type',
            MEAN=''*RESPONSE=''*(WILSON_LEN='Wilson' WALD_LEN='Wald')*FORMAT=5.1
            /ROW=FLOAT MISSTEXT='*';
            TITLE 'Interval lengths';
RUN;


/* FREQUENCIES IN POPULATION */
PROC SORT DATA=DATA.POPULATION;
            BY COUNTY;
RUN;
PROC MEANS DATA=DATA.POPULATION NOPRINT;
            VAR
            ALL_OFF_REP_COUNTY
            ROBBERY_REP_COUNTY
            ASSAULT_REP_COUNTY
            SEXUAL_REP_COUNTY
            THREAT_REP_COUNTY
            FRAUD_REP_COUNTY
            HARASS_REP_COUNTY
            ANY_REP_COUNTY
            ROBBERY
            ASSAULT
            SEVERE_ASSAULT
            SEXUAL
            THREAT
            FRAUD
            HARASS
            ANY
            SEX
            CIVIL_STATUS
            AGECLASS
            INCOMECLASS
            COUNTY;
            BY COUNTY;
            OUTPUT OUT=DATA.COUNTY_TRUE_VALUES
            MEAN(ALL_OFF_REP_COUNTY ) = ALL_OFF_REP_COUNTY
            MEAN(ROBBERY_REP_COUNTY) = ROBBERY_REP_COUNTY
            MEAN(ASSAULT_REP_COUNTY) = ASSAULT_REP_COUNTY
            MEAN(SEXUAL_REP_COUNTY) = SEXUAL_REP_COUNTY
            MEAN(THREAT_REP_COUNTY) = THREAT_REP_COUNTY
            MEAN(FRAUD_REP_COUNTY) = FRAUD_REP_COUNTY
            MEAN(HARASS_REP_COUNTY) = HARASS_REP_COUNTY
            MEAN(ANY_REP_COUNTY) = ANY_REP_COUNTY
            MEAN(ROBBERY ) = ROBBERY
            MEAN(ASSAULT) = ASSAULT
            MEAN(SEVERE_ASSAULT) = SEVERE_ASSAULT
            MEAN(SEXUAL) = SEXUAL
            MEAN(THREAT) = THREAT
            MEAN(FRAUD) = FRAUD
            MEAN(HARASS) = HARASS
            MEAN(ANY) = ANY
            ;
RUN;
PROC MEANS DATA=DATA.COUNTY_TRUE_VALUES;
```

```
            VAR
            ALL_OFF_REP_COUNTY
            ROBBERY_REP_COUNTY
            ASSAULT_REP_COUNTY
            SEXUAL_REP_COUNTY
            THREAT_REP_COUNTY
            FRAUD_REP_COUNTY
            HARASS_REP_COUNTY
            ANY_REP_COUNTY
            ROBBERY
            ASSAULT
            SEVERE_ASSAULT
            SEXUAL
            THREAT
            FRAUD
            HARASS
            ANY;
RUN;
PROC SORT DATA=DATA.POPULATION;
            BY MUNICIPALITY;
RUN;
PROC MEANS DATA=DATA.POPULATION NOPRINT;
            VAR
            ALL_OFF_REP_MUNICIPALITY
            ROBBERY_REP_MUNICIPALITY
            ASSAULT_REP_MUNICIPALITY
            SEXUAL_REP_MUNICIPALITY
            THREAT_REP_MUNICIPALITY
            FRAUD_REP_MUNICIPALITY
            HARASS_REP_MUNICIPALITY
            ANY_REP_MUNICIPALITY
            ROBBERY
            ASSAULT
            SEVERE_ASSAULT
            SEXUAL
            THREAT
            FRAUD
            HARASS
            ANY
            SEX
            CIVIL_STATUS
            AGECLASS
            INCOMECLASS
            MUNICIPALITY;
            BY MUNICIPALITY;
            OUTPUT OUT=DATA.MUNICIPALITY_TRUE_VALUES
            MEAN(ALL_OFF_REP_MUNICIPALITY ) = ALL_OFF_REP_MUNICIPALITY
            MEAN(ROBBERY_REP_MUNICIPALITY) = ROBBERY_REP_MUNICIPALITY
            MEAN(ASSAULT_REP_MUNICIPALITY) = ASSAULT_REP_MUNICIPALITY
            MEAN(SEXUAL_REP_MUNICIPALITY) = SEXUAL_REP_MUNICIPALITY
            MEAN(THREAT_REP_MUNICIPALITY) = THREAT_REP_MUNICIPALITY
            MEAN(FRAUD_REP_MUNICIPALITY) = FRAUD_REP_MUNICIPALITY
            MEAN(HARASS_REP_MUNICIPALITY) = HARASS_REP_MUNICIPALITY
            MEAN(ANY_REP_MUNICIPALITY) = ANY_REP_MUNICIPALITY
            MEAN(ROBBERY ) = ROBBERY
            MEAN(ASSAULT) = ASSAULT
            MEAN(SEVERE_ASSAULT) = SEVERE_ASSAULT
            MEAN(SEXUAL) = SEXUAL
            MEAN(THREAT) = THREAT
            MEAN(FRAUD) = FRAUD
            MEAN(HARASS) = HARASS
            MEAN(ANY) = ANY
            ;
RUN;
PROC MEANS DATA=DATA.MUNICIPALITY_TRUE_VALUES;
            VAR
            ALL_OFF_REP_MUNICIPALITY
            ROBBERY_REP_MUNICIPALITY
            ASSAULT_REP_MUNICIPALITY
            SEXUAL_REP_MUNICIPALITY
            THREAT_REP_MUNICIPALITY
            FRAUD_REP_MUNICIPALITY
            HARASS_REP_MUNICIPALITY
            ANY_REP_MUNICIPALITY
            ROBBERY
            ASSAULT
            SEVERE_ASSAULT
```

```
            SEXUAL
            THREAT
            FRAUD
            HARASS
            ANY;
RUN;
PROC MEANS DATA=DATA.POPULATION;
            VAR
            ROBBERY_REPORT
            ASSAULT_REPORT
            SEVERE_ASSAULT_REPORT
            SEXUAL_REPORT
            THREAT_REPORT
            FRAUD_REPORT
            HARASS_REPORT
            ANY_REPORT
            ROBBERY
            ASSAULT
            SEVERE_ASSAULT
            SEXUAL
            THREAT
            FRAUD
            HARASS
            ANY;
RUN;
PROC FREQ DATA=DATA.POPULATION;
            table (SEX CIVIL_STATUS AGECLASS INCOMECLASS NON_NORDIC URBAN) / NOCOL NOCUM;
RUN;
```