Bachelor's project

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Andreas Matre

Linear Regression for Survey Data

Bachelor's project in Mathematical Sciences
Supervisor: Geir-Arne Fuglstad
June 2020

NTNU
Norwegian University of Science and Technology

Andreas Matre

Linear Regression for Survey Data

Bachelor's project in Mathematical Sciences
Supervisor: Geir-Arne Fuglstad
June 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Kunnskap for en bedre verden

Abstract

This Bachelor's thesis is submitted for the course MA2002 at NTNU which is 15 credits over one semester.

In this thesis we discuss how to do linear regression when the data is collected using a complex sampling design. We use a different paradigm from classical regression, where we assume an infinite population and that the response observed for each individual is random. Here, we instead acknowledge that the population is finite and assume the value of each individual is fixed and the randomness arises from which individuals are included in the sample. The major issues are accounting for different sampling designs to prevent bias and incorrect uncertainty estimates.

We explain three different sampling techniques: the first sampling technique we look at is a Simple Random Sample. First, we choose the size of the sample. Then we let each possible subset of the population, of that sample size, have the same probability of being chosen as the sample. A Simple Random Sample has the advantage that all the sampled units are independent. The second sampling technique we look at is stratification. Here we split the population into a partition and sample independently from each subset. This allows us to get independent regression lines from each subset. The third sampling technique we look at is clustering. Here we again split the population into a partition, but instead of sampling from all subsets of the partition we instead sample only from some of the subsets, chosen by taking a sample of the subsets. Clustering is used to reduce costs when doing surveys. Often the clusters are geographical areas, which means that sampling only inside some subsets allow us to save travel time. When performing large surveys, these techniques are usually combined into what is called a complex survey. For example, by first doing stratification on the whole population and then using clustering inside each subset.

If the sampling units inside the strata are similar, then stratification will reduce the uncertainty compared to a SRS of the same size. With clustering, however, we usually get larger uncertainty, as units inside clusters are usually more similar than units across clusters. This causes the sample to carry less information than a non clustered sample. This leads to hypothesis tests regarding the regression line having less power and the prediction intervals to become larger. The non-linear nature of the regression coefficients means that estimating their variance becomes complicated. We therefore show an approximation technique called linearization.

Denne bacheloroppgaven er en del av emnet MA2002 på NTNU, som er 15 studiepoeng over ett semester.

I denne oppgaven diskuterer vi hvordan man gjør linear regresjon når et komplekst utvalgsdesign er brukt for å samle inn dataene. Vi bruker et annet paradigme enn i klassisk regresjon hvor vi antar en uendelig populasjon og at hver observerte verdi er tilfeldig. Her aksepterer vi istedenfor at populasjonen er endelig og antar at hver respons er fastsatt og det tilfeldige kommer fra hvilke individer som er med i utvalget. De største problemene er å ta hensyn til de forskjellige utvalgsdesignene for å unngå bias i resultatene, samt feil variansestimater.

Vi skal gå gjennom tre ulike utvalgsteknikker: Den første utvalgsteknikken vi ser på er en SRS. Her velger man først størrelsen på utvalget. Så lar man hvert mulige utvalg, med den valgte størrelsen av populasjonen, ha samme sannsynlighet for å bli valgt. En SRS har fordelen at alle individer i utvalget er uavhengige. Den andre utvalgsteknikken vi ser på er stratifisering. Her deler vi populasjonen i en partisjon og gjør et uavhengig utvalg fra hver delmengde. Dette gjør at vi kan lage uavhengige regresjonslinjer for hver delmengde. Den tredje utvalgsmetoden vi ser på er klyngeutvalg. Her deler vi igjen populasjonen inn i en partisjon, men istedenfor å gjøre et utvalg fra alle delmengdene gjør vi bare et utvalg fra noen av dem, valgt ved å gjøre et utvalg av delmengdene. Klyngeutvalg er brukt for å redusere kostnadene ved å gjøre utvalg. Ofte er delmengdene geografiske områder, som betyr at å bare gjøre utvalg innen noen delmengder sparer reisetid. Når man utfører store

undersøkelser er ofte disse metodene kombinert til det som kalles en kompleks undersøkelse. For eksempel gjør man ofte først stratifisering på hele populasjonen, og så bruker man klyngeutvalg innen hver delmengde.

Hvis individene innen delmengdene i stratifisering er like så vil stratifisering redusere usikkerheten i estimatene sammenlignet med en SRS av samme størrelse. Ved klyngeutvalg, derimot, vil vi som regel få større usikkerheter i estimatene ettersom individene inne i delmengdene ofte er mer like enn individer på tvers av delmengdene. Dette gjør at utvalget inneholder mindre informasjon om populasjonen enn et ikke-klusteret utvalg. Dette fører til at hypotesetester angående regresjonslinjen for lavere styrke og at prediksjonsintervallene blir større. Det at regresjonskoeffisientene er ikke-lineære uttrykk gjør at variansestimasjon er komplisert. Vi viser derfor en approksimasjonsmetode som heter linearisering.

Contents

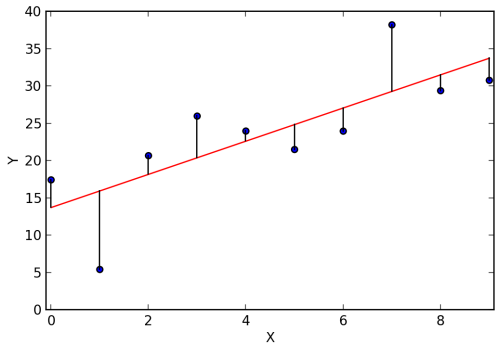


Figure 1: Illustration of residuals. The red line is a regression line. The distances between the points and the line are illustrated by the vertical black lines. https://upload.wikimedia.org/wikipedia/commons/e/ed/Residuals_for_Linear_Regression_Fit.png

1 Introduction

Linear regression is a useful tool to describe relationships between variables. When we want to investigate possible associations, linear regression is one of the most common tools we use. It can be used to show correlation between variables, and in some cases, where experiments are designed very carefully, it can even suggest causation. It can, for example, be used in the social sciences to try to show a relationship between income and age of death. In simple linear regression, we want to determine the line that best describes the relationship between the predictor and the response. Since the relationship is usually not exactly linear, there is almost always some noise around the line, which gives us residuals, the vertical vectors that are between the points and the linear line. The residuals are illustrated in Figure 1. If the relationship was exactly linear we would just need to sample two individuals, with different predictor values, to find the regression line. When doing linear regression, the goal is to find the line that minimizes the sum of the residuals squared.

Since we are doing measurements on a real population, it is also finite. If we knew the measurements for the whole population, there would be no uncertainty. In that case we could just calculate whatever quantity we wanted, including the regression intercept and slope. When we generalize from a sample to the whole population, however, we get uncertainty because we can never know the values of the individuals not in the sample. There are different approaches of modeling this uncertainty, and which method we should choose depends on what we know about the data. For example, which methods were used in collecting the data? What is the probability of each individual being included in the sample? Are the individuals sampled independently?

One approach to model the data is called the model based approach and is based on the fact that we usually sample from large populations. We assume that the relationship between the predictor and the response is split in two: a linear deterministic part and a stochastic part representing the error, which can be described by a continuous distribution. We assume that the population is large enough that the probability of sampling the same unit again if we take a new sample is negligible. This allows us to assume that when we sample an unit, the residual can be thought of as being sampled from the distribution. We can use the structure from this assumption to model the uncertainty of the coefficients and predictions. The only thing we need to know about the data collection method is the fact that the individuals were sampled

independently. In the model based approach, we assume there exists a line representing the deterministic part of the relationshiop between the predictor and response. Our goal is to use a sample of the population to get a estimate of the true line. The assumption that the probability of sampling the same unit again is negligible if we resample is often not realistic, however, nor the fact that the residuals follow a distribution. This motivates another approach to model the data.

A, perhaps, more realistic approach to modeling the data is to assume that the responses are fixed, in which case we can not pretend to be drawing values from a distribution. To model the uncertainty in this approach, we look at the sampling probabilities, i.e, the probability of each individual being included in the sample. Or, said another way, the random part of the model is who is sampled as opposed to the values of the sampled units. Knowing the sampling probabilities is enough to estimate the regression intercept and slope, but to estimate the uncertainties, we also need information regarding how the sample was collected. Using this type of information, and the fact the the random part is now who is in the sample, is called a design based approach. The design based approach requires more information about the sampling scheme than the model based approach. The estimated variances with a design based approach are often larger than when using a model based approach. This is because the assumption that the residuals are drawn from a specific distribution gives a lot of extra structure. In the design based approach our goal in regression is to find the line minimizing the residual sum of squares for the whole population.

The population we are sampling from could be heterogenuous, i.e, different parts of the population have very different response values. Say, for example, we want to find the mean income of residents of Oslo. When sampling, we could risk getting only people living in the west part of Oslo, which is the most wealthy area of the city. This would result in the estimate being much higher than the true mean income of the residents of Oslo. To fix this, we could split the city population into subsets; one for each city district. We can then sample from each of these districts independently. Doing that, we are guaranteed to get a sample including people from different parts of the city, and therefore more likely to get correct estimates. This can dramatically decrease the uncertainty of the estimates when the subsets are chosen smartly. Doing this also allows us to make separate estimates of the quantity of interest for different parts of the population, i.e, we could make a separate estimate of the mean income for each city district in addition to the estimate for the city as a whole. This method is called stratification.

Another potential problem is that researchers are often on a limited budget, and it can be expensive to sample randomly from the whole population. Say, for example, that each sampled individual requires a visit from an interviewer. If the sample is spread randomly in the country, this can get expensive as the interviewers would need to use much time to travel. One often solves this by splitting the country into geographic parts, for example municipalities, and then randomly choose some municipalities to sample people from. One then pick independent samples from the chosen municipalities. This has the advantage that it saves travel time, and, therefore, makes it cheaper to conduct the survey. The problem with this approach, however, is that the sampled units become dependent since we sample from only some municipalities. This results in larger variances. This can be remedied by the fact that we can sample more individuals inside just some municipalities for the same cost as having a smaller sample from the whole population. This is called clustering.

We often combine stratification and clustering, which results in complicated sampling schemes where we, for example, first cluster on municipalities, then within each municipality we may, for example, split the population by age and sample from all the age groups. Because this gets complicated, it is often difficult to find explicit formulas for the variance, and we usually have to estimate the variance instead. The fact that we have a finite population, also influences the variance, since we get something called the finite population correction. This is a factor $1 - \frac{n}{N}$,

where n is the size of the sample and N is the size of the population. This factor takes into account the fact that as we get a larger and larger sample we can learn everything about the population, and therefore the variance goes to zero.

This thesis shows how to fit a linear regression model when data is collected through a complex survey design, i.e, a survey including unequal sampling probabilities, stratification and clustering. To do analysis in these cases weights are used. Each observation gets a weight value which can be interpreted as the number of individuals in the population that observation represents. So a unit with a small chance of being included in the sample would have a larger weight than a unit with a high chance of being included. In classical regression each individual in the population has the same chance of being included so each observation therefore has the same weight.

We start with an example illustrating what can go wrong if the design used in collecting the data is not taken into account.

Example 1. *We use a dataset from a study of the relationship between the length of a persons left middle finger and their height. The researcher oversampled short people and undersampled tall people. The dataset contains 200 samples, each containing the length of the persons left middle finger (cm), their height (cm) and the probability that they would be chosen for the sample.*

To illustrate the difference, the top left plot in Figure 2 shows a random sample where every person had an equal chance of being included. While the top right plot in Figure 2 shows the sample where short people had a higher chance of being included than tall people. The observations in the right panel are much more concentrated in the bottom. This means that fitting a linear regression model to the unequal probabilities sample will result in the slope being smaller than it should be.

The bottom row of Figure 2 illustrates the difference between classical regression, assuming a normal distribution for the residuals, and regression taking the sampling probabilities into account. The bottom left plot shows regressions based on a sample where every person has the same probability of being included. We can see only one regression line in this plot, this is because both regression lines are exactly equal in this case. In the bottom right plot of Figure 2, we see there are two different lines: the red line is from classical regression while the blue line is from regression taking sampling probabilities into account. The shaded areas represent 95% prediction intervals. We can see that the red line seems to fit the sample better than the blue one which seems too steep. This is because the blue line takes the oversampling of people with short fingers into account. The blue line is almost identical to the regression line in the bottom left plot, which has no bias from unequal sampling probabilities.

These differences show that it is important to take the design of the survey into account when analyzing the data, because classical linear regression can give misleading estimates.

The rest of the thesis focuses on a dataset on the performance of students in schools in California. The dataset has data on all 6194 schools that have more than 100 students in California. The data collected includes: API scores in 1999 and 2000, which level of school it is (elementary, middle, high), name of school, location of school, percentage of students tested at the school, API targets, economic factors for students at the school, class sizes, information of education of parents and qualification of teachers. Each school gets a API score, which is a metric quantifying the academic performance of the students at the school. We will use this data as the population we will sample from, and we will take different kinds of samples to illustrate the concepts introduces in this thesis. See ((Lumley, 2010, Section 1.2)) for more details on the dataset.

In this thesis we will use the "survey" package in R to make computations, Lumley ((2020)).

Section 2 will give a overview of model based simple linear regression, which is assumed known and only a brief repetition will be given. Section 3 will explain how to do linear regression in

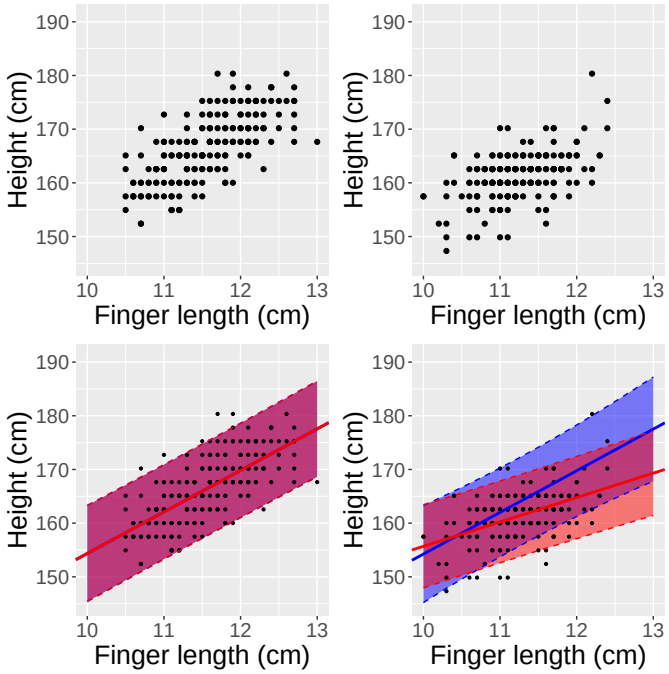


Figure 2: The top row shows the two samples from the finger length versus height dataset. The top left plot shows a sample where everyone has the same chance of being included in the sample. The top right plot shows a sample where people with short fingers are oversampled. The bottom row shows estimated regression lines based on the samples. Each bottom plot shows estimated regression lines for the sample in the plot above. The red lines are estimated regression lines using classical regression while the blue lines are estimated regression lines taking the probabilities of being included in the sample into account. The shaded areas represent the 95% prediction intervals. Observe that in the bottom left plot the two approaches agree while in the bottom right plot, they do not agree.

the context of finite populations using survey statistics. Section 3.1 explains the general theory behind regression in finite population, while Section 3.2 will show how to do linear regression when we have a Simple Random Sample. Section 4 talks about estimating quantities with more complex survey designs: Section 4.1 will go through estimation when we have a sample using stratification. Section 4.2 will explain estimation when we have a sample using clustering and Section 4.3 will show how to do estimation when we have surveys combining stratification and clustering into a complex survey. Section 5 is about estimating the variance of non-linear expressions, while Section 6 is the discussion.

2 Classical simple linear regression

In classical simple linear regression, each observation i consists of a response variable, y_i, and a predictor, x_i for $i = 1, ..., n$, where n is the number of observations. The relationship between the response and the covariates is assumed to be

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, \ldots, n,$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are stochastic variables describing the error. The intercept, β_0, and the slope, β_1, are constants that describe the deterministic part of this relationship. We do not know the values of β_0 and β_1, so we have to estimate them from observed data points.

We call the data points $(x_1, y_1), (x_2, y_2)...(x_n, y_n)$. To estimate the deterministic part of the relationship we want to find estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the Residual Sum of Squared (RSS). The RSS is defined by

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 .$$

RSS can be geometrically thought of as the sum of the squared distances of the data points in the sample from the regression line, see Figure 1. We want the line to follow the linear trend in the sample as closely as possible, and minimizing the RSS is one way to find a close line.

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the RSS are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{1}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \tag{2}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. ((Larsen and Marx, 2012, Chapter 11))

Assume the following conditions:

1. $\mathrm{E}(\epsilon_i) = 0 \ \forall \ i = 1, 2, ..., n$
2. $\mathrm{Var}(\epsilon_i) = \sigma^2 \ \forall \ i = 1, 2, ..., n$
3. All the ϵ are independent of any predictor or observation number.
4. All ϵ_1, ϵ_2, ..., ϵ_n are independent of each other,

Here item 1 means that data points should on average lie on the line.
Item 2 means that the data points should have the same variance around the regression line.

Item 3 and 4 means that there should be no pattern in whether the data points is over or under the regression line and on the vertical Euclidean distance from the regression line.

Under these conditions we get some useful properties, including that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1. In addition we have unbiased estimates for the variance of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\widehat{\text{Var}}\left(\hat{\beta}_0\right) = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$\widehat{\text{Var}}\left(\hat{\beta}_1\right) = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2},$$

where $\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$ is an unbiased estimator for the unknown σ^2, ((Larsen and Marx, 2012, Chapter 11)).

This is based on the fact that we assume the response is random, i.e, that we can think of the residuals as coming from some stochastic distribution. We will now consider what happens if we instead assume that the responses are fixed and the randomness comes from how we select individuals instead. This approach is called survey statistics

3 Regression in the context of finite populations

3.1 General results

Now consider the case where the population is finite, i.e, the population consists of $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, where the x_i's are the covariates we use to predict y_i and N is the size of the population. The goal of linear regression in this case is that we want to find the line, $y = B_0 + B_1 x$, that best describes the relationship between x_i and y_i in this population. We define the best line as the one that minimizes RSS $= \sum_{i=1}^{N}(y_i - B_0 - B_1 x_i)^2$. Minimizing the RSS means that we minimize the squared Euclidian distance of each point in the population from the line.

This differs from model based linear regression where we want to estimate the deterministic part of the relationship between x_i and y_i, but we can never get an exact answer as the estimates will differ when they are based on different samples, no matter how large the samples are. Here, however, it is possible to find B_0 and B_1, because we can simply sample the whole population, even if that typically is not possible in practice.

If we, however, knew the whole population, we could just compute B_0 and B_1 using the same formulas as in Section 2. For our purposes we will rewrite Equations 1 and 2, so that they are expressed by totals of the population. They therefore become

$$B_0 = \frac{1}{N}\left(t_y - \frac{t_{xy}t_x - \frac{1}{N}t_y t_x^2}{t_{x^2} - \frac{1}{N}t_x^2},\right) \tag{3}$$

$$B_1 = \frac{t_{xy} - \frac{1}{N}t_y t_x}{t_{x^2} - \frac{1}{N}t_x^2}, \tag{4}$$

where $t_x = \sum_{i=1}^{N} x_i$, $t_y = \sum_{i=1}^{N} y_i, t_{x^2} = \sum_{i=1}^{N} x_i^2$ and $t_{xy} = \sum_{i=1}^{N} x_i y_i$, ((Lohr, 2009, Chapter 11)).

The case where we know the whole population is not realistic. We are interested in the case where we have to sample from the population to estimate B_0 and B_1. To do that we have to introduce some terms.

Definition 1. *A **sampling unit** is one "element" we can sample. The **sampling population**, or universe, $U = \{1, 2, 3, ..., N\}$, is a finite set containing all the sampling units we can sample.*

In the case of our API dataset, our sampling units are schools, while the sampling population is all the schools in California having more than 100 students.

Definition 2. *A **sampling frame** is a list of sampling units that one uses to draw a sample.*

The sampling frame would be all schools in California that the researchers know about and that the researchers think have more than 100 students. Ideally the sampling frame and the sampling population would be the same, but that is not always the case. When taking different types of samples from the API dataset, the sampling population and the sampling frame are equal, since we have a table of all the data and we will just choose rows from that table for our samples. But there are cases where this is not the case. An example of this could be if we were doing a political survey to try to predict who will win the next election. In that case our sampling population, who we are interested in information about, would be everyone who are going to vote in the next election. It is, however, impossible to get a list of them, so we instead have to use some other part of the population we have information about as our sampling frame. We might have a list of all who voted last election, which might be a good approximation of those who are going to vote this election, but then we would miss out on all the new eligible voters and people who might have decided to vote this election but didn't do it the last one. Choosing the correct sampling frame to match the target sampling population is difficult and if they do not match it may influence the results.

We will assign an unique integer index to each sampling unit and list them in an arbitrary order. This simplifies notation.

Definition 3. *A **sample**, $S \subseteq U$, is a subset of the sampling frame. This is the data we will analyse to learn about the sampling population. A **probability sample** is a sample where the sampling units included are chosen randomly. The **sampling probability** of a sampling unit is the probability that a specific sampling unit will be included in the sample.*

If we let $\hat{t}_x, \hat{t}_y, \hat{t}_{x^2}, \hat{t}_{xy}, \hat{N}$ be estimators for $t_x, t_y, t_{x^2}, t_{xy}, N$, respectively, we get estimators for B_0 and B_1 by replacing quantities by estimated quantities in Equations 4 and 3 respectively

$$\hat{B}_1 = \frac{\hat{t}_{xy} - \frac{1}{\hat{N}}\hat{t}_y\hat{t}_x}{\hat{t}_{x^2} - \frac{1}{\hat{N}}\hat{t}_x^2}$$

$$\hat{B}_0 = \frac{1}{\hat{N}}\left(\hat{t}_y - \frac{\hat{t}_{xy}\hat{t}_x - \frac{1}{\hat{N}}\hat{t}_y\hat{t}_x^2}{\hat{t}_{x^2} - \frac{1}{\hat{N}}\hat{t}_x^2}\right) = \frac{\hat{t}_y}{\hat{N}} - \hat{B}_1\frac{\hat{t}_x}{\hat{N}}$$

Since $\hat{B}_0$ and $\hat{B}_1$ are non linear expressions of dependent statistics, deriving exact expressions for the variances is complicated. We, therefore, often have to settle with having estimates of the variances instead. There are several ways to do so, but a common one, and the one we use in this thesis, is linearization. Linearization takes a non-linear expression of stochastic variables we want to do inference about and uses the first two terms of the Taylor expansion to make it linear. One can show that

$$\mathrm{Var}(\hat{B}_1) \approx \frac{\widehat{\mathrm{Var}}\left(\sum_{i\in S} w_i q_i\right)}{\left(\sum_{i\in S} w_i x_i^2 - \frac{\left(\sum_{i\in S} w_i x_i\right)^2}{\sum_{i\in S} w_i}\right)^2}$$

where $q_i = (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}})$. $\widehat{\text{Var}}\left(\sum_{i \in S} w_i q_i\right)$ is easier to work with as $\sum_{i \in S} w_i q_i$ estimates a total. See ((Lohr, 2009, Chapter 11.2.2)) for details on derivation. We describe linearization in more detail in Section 5.

To calculate the estimated regression coefficients we need to know how to create estimates of the totals we need for different sampling designs.

3.2 Simple random sample

The simplest probability sample is the Simple Random Sample (SRS). A sample of size $n \leq N$ is an SRS if every subset $S \subseteq U$ has the same probability of being chosen. If, for example, $U = \{1, 2, 3, 4\}$ and we want a sample, S, of size 3, then there are $\binom{4}{3} = 4$ possible samples: $S_1 = \{1, 2, 3\}$, $S_2 = \{1, 2, 4\}$, $S_3 = \{1, 3, 4\}$ and $S_4 = \{2, 3, 4\}$.

For this to be a SRS each of these subsets need to have the same probability of being chosen, i.e, $P(S_1) = P(S_2) = P(S_3) = P(S_4) = 0.25$. A consequence of having a SRS is that all the sampling probabilities are equal, $P(1 \in S) = P(2 \in S) = P(3 \in S) = P(4 \in S) = 0.75$. But having equal sampling probabilities is not sufficient for the sample to be an SRS. Look, for example, at this case: Assume we want a sample of size 2 from a population of size 4, and that $P(\{1, 3\}) = 0.5$ and $P(\{2, 4\}) = 0.5$ while the probabilities of all the other possible samples are 0. Then $P(1 \in S) = P(2 \in S) = P(3 \in S) = P(4 \in S) = 0.5$ but this is not a SRS since all possible subsets of size 2 do not have equal probability of being chosen. This is actually an example of a cluster sample which is discussed in Section 4.2.

We need estimates of several different totals of the population to estimate the regression coefficients, see Equations 3 and 4. We need the total of, among others, the y_i's, the x_i's and the $x_i y_i$'s, to calculate the estimates. We will do inference on the total of the y_i's in this thesis. The other totals are equivalent and not shown.

Let $t_y = \sum_{i=1}^{N} y_i$ be the value we want to estimate. The natural estimator of this total, if we sample n elements, would be $\hat{t}_y = \frac{N}{n} \sum_{i \in S} y_i$ where we take the average of the values in our sample and then scale it up to the whole population. It can be shown, using indicator variables and sampling probabilities, that $\hat{t}_y$ is an unbiased estimator for t_y.

It can also be shown that the variance of the estimator is of the form

$$\text{Var}\left(\hat{t}_y\right) = \frac{N^2}{n\,(N-1)} \left(1 - \frac{n}{N}\right) \sum_{i=1}^{N} (y_i - \bar{y})^2$$

where $\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$ is the variance of the whole population, see ((Lohr, 2009, Chapter 2)). We do not, however, know the population variance, as that would require us to know the y values for the whole population, U. Instead we estimate the population variance by the unbiased estimator $\frac{1}{n-1} \sum_{i \in S} \left(y_i - \hat{\bar{y}}\right)^2$, where $\hat{\bar{y}} = \frac{1}{n} \sum_{i \in S} y_i$, which gives us the estimate of $\text{Var}(\hat{t}_y)$

$$\widehat{\text{Var}(\hat{t}_y)} = \frac{N^2}{(n-1)\,n} \left(1 - \frac{n}{N}\right) \sum_{i \in S} \left(y_i - \hat{\bar{y}}\right)^2$$

Since $\frac{1}{n-1} \sum_{i \in S} \left(y_i - \hat{\bar{y}}\right)^2$ is an unbiased estimator of $\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$, we have that $\widehat{\text{Var}(\hat{t}_y)}$ is an unbiased estimator of the variance of $\hat{t}_y$.

We see that all these estimators for totals and means are the same as in the model based case. Therefore, the estimates for the coefficients in the regression model are also the same. The variance of the estimators for the coefficients are different however. The factor $\left(1 - \frac{n}{N}\right)$ in the variances is what differs in the variance estimate compared to the model based one. It is called the **finite population correction (fpc)** and comes from the fact that we are sampling without

replacement from a finite population. For an intuitive explanation of the fpc, consider we take a sample of size 10. If the population size is just 15 we would expect to have a lot more information about the whole population than if the population size was large. The fpc also makes sure that the variance is 0 if we sample the whole population. Note, however, that we will not have a deterministic model, even if we sample the whole population. This is because the data points lay around the line and not on it, so there will always be uncertainty when predicting.

Using linearization we get this estimate for the variance of $\hat{B}_1$ when the sample is from an SRS, see ((Lohr, 2009, Chapter 11.2))

$$\widehat{\text{Var}}(\hat{B}_1) = \left(1 - \frac{n}{N}\right) \frac{n}{n-1} \frac{\sum_{i \in S} (x_i - \bar{x})^2 \left(y_i - \hat{B}_0 - \hat{B}_1 x_i\right)^2}{\left(\sum_{i \in S} (x_i - \bar{x})^2\right)^2} \tag{5}$$

which we can compare to the estimate of the variance for $\hat{\beta}_1$ in the model based case

$$\widehat{\text{Var}}\left(\hat{\beta}_1\right) = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{(n-2) \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{6}$$

where $\bar{x} = \frac{1}{n} \sum_{i \in S} x_i$.

One important difference between the variance estimates is that in Equation 5 we have the fpc, which comes from the fact that we sample from a finite population.

When doing surveys, SRSs are often not used. Instead one uses more complex survey techniques which can make the estimated quantities have smaller uncertainties and make the survey cheaper to perform. We will therefore now look at different sampling designs that are often used. Not taking the sampling design into account can lead to a bias in the results and too small or too large variances. We will focus on estimating totals of the population, as the regression coefficients are expressed as functions of totals in Equations 3 and 4.

4 Accounting for survey design

4.1 Stratification

In stratification we split the sampling frame into a partition, i.e, H non-overlapping subsets that together comprise the whole sampling frame. These subsets are called **strata**. We let each stratum have $N_i,\ i = 1, \ldots, H$, elements. Thus $N_1 + N_2 + \cdots + N_H = N$. When sampling we independently draw samples from each stratum, $S_1, S_2, \ldots, S_H$, with $n_i,\ i = 1, \ldots, H$, elements. When estimating a total we can first estimate the total of each stratum and then add these estimated totals to get an estimate of the population total. If we let $t_{y,h}$ be the total in stratum h, for $h = 1, 2, \ldots, H$, we get $\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{y,h}$, where we can use different sampling schemes to estimate each $t_{y,h}$. If we let the sample of each stratum be a simple SRS, we get $\hat{t}_y = \sum_{h=1}^{H} \sum_{i \in S_h} \frac{N_h}{n_h} y_i$. Since the estimate for each stratum total is unbiased (see Section 3.2), the estimate of the population total is unbiased.

Since we often sample differently in the different strata, the individuals sampled from the different stratum usually have different sampling probabilities. This means that different sampling units should be weighted differently when making estimates, as illustrated in Example 2.

Example 2. *Suppose a population is divided into two strata, each with a subpopulation of* 1000 *individuals. Let one subpopulation be values with mean* 0 *and variance* 1 *and one subpopulation*

be values with mean 10 *and variance* 1*, called A and B respectively. Our goal is to estimate the sum of the values.*

The true sum of the population values is 10058.59*. If we sample the same proportion of values from A and B we can estimate the sum the same way as in a SRS. Let* $S_{A,100}$ *and* $S_{B,100}$ *be SRS's of size* 100 *from A and B respectively. Then an unbiased estimate of the total is* $\sum_{i \in S_{A,100} \cup S_{B,100}} \frac{2000}{200} y_i = 10104.81$*. The bias is only* 46.22 *which is only* 0.46% *of the value.*

If the proportion of units sampled from A is different from the proportion of units sampled from B, however, we can't use this simple estimate. Let $S_{A,50}$ *be a SRS of size* 50 *from A. If we use the same SRS estimate again we get* $\sum_{i \in S_{A,50} \cup S_{B,100}} \frac{2000}{150} y_i = 13560.02$*. The bias is here* 3501.425 *which is* 34.8% *of the value. This is a much larger error. This is because we have more samples from B than A, and B has a higher mean than A. Since each sampled value is counted the same this causes the estimate to become too large. To fix this we need to let the sampled values from A count, or weigh, more than the ones from B. This is because the values in* $S_{A,50}$ *have to represent the same size population as the values in* $S_{B,100}$*, but there are only half as many values in* $S_{A,50}$ *as in* $S_{B,100}$*. Therefore, each value in* $S_{A,50}$ *has to count two times as much as the values in* $S_{B,100}$*.*

Doing this the estimate becomes $\sum_{i \in S_{A,50}} \frac{1000}{50} y_i + \sum_{i \in S_{B,100}} \frac{1000}{100} = 10174.61$ *which has a bias of* 116 *which is* 1.15%*, a huge improvement over the* 34.8% *bias where we did not take the different sampling probabilities into account.*

This illustrates how important it is to take the sampling probabilities into account. To do this we usually give each sampled unit a weight value,

Definition 4. *The* ***weight*** *of a sampling unit is the inverse of the sampling probability of the sampling unit. We denote the weight of unit* i *as* w_i*.*

An intuitive way to interpret weights is that the weight of an observation is how many sampling units in the population they represent. If we sample few units from a large strata, each of these sampled units represents many more individuals than in a strata where we sample almost the whole subpopulation. The name weight makes sense, as an observation representing many unobserved units should "weigh" more when estimating values than observations representing few unobserved units.

Specifically, in Example 2, the 50 sampled units from $S_{A,50}$ represent a population of 1000, so each sampled unit represents 20 units including itself. This is opposed to the 100 sampled units in $S_{B,100}$ which also represent a population of 1000. Here each sampled unit only represents 10 units including itself.

Using weights, we can rewrite the estimate of t_y in a more general form, which is valid for any sampling scheme, $\hat{t}_y = \sum_{i \in S} w_i y_i$, where the full sample is $S = S_1 \cup S_2 \cup \cdots \cup S_H$. This is a convenient way to calculate estimates of more complicated surveys, as one only needs to calculate the weights once, and then one can use them to estimate many different quantities. This also works for an SRS as it can be shown that the sampling probability of a sampling unit in a SRS is $\frac{n}{N}$. This means that the weight of each sampling unit is $\frac{N}{n}$. Therefore the estimator using weights is the same as the one introduced in Section 3.2, $\hat{t}_y = \sum_{i \in S} \frac{N}{n} y_i$.

Since the samples from the different strata are independent, the variance of the estimator is also easy to calculate, $\text{Var}(\hat{t}) = \text{Var}\left(\sum_{h=1}^{H} \hat{t}_{y,h}\right) = \sum_{h=1}^{H} \text{Var}\left(\hat{t}_{y,h}\right)$. This means that to minimize the variance of $\hat{t}$ we should choose the strata such that the internal variance in each stratum is as small as possible.

Stratification is used for several reasons: making it possible to analyze subpopulations individually, making sure subgroups are included in the sample, and reducing uncertainty.

	Elementary schools	Middle schools	High schools
Estimates (95% confidence interval)	169.1 (136.9, 201.3)	145 (121, 169)	143.8 (119.7, 167.9)
True value	144.5	157.7	152.2

Table 1: Table of slope coefficients for the different school levels. The first row has estimates from the sample, with 95% confidnce intervals in parenthesis. The second row has the values, as we know the whole population.

If we have a population with several different interesting subgroups, it can be useful to let each of these subgroups be their own strata. This will allow us to create a seperate regression line for each stratum, which will allow us to analyze them by themselves. We can then compare the slopes to see if there is a difference in the relationship between the response and predictor for the different strata. We could of course make regression lines for different subgroups after doing a sample without stratification, but then we would have no guarantee that each of the subgroups would have enough samples to be able to make a useful regression line. Using stratification, we can choose how many samples we want from each subgroup.

Example 3. *Suppose that we want to investigate the relationship between the API score of a school and the average level of education of the student's parents. We might be interested in knowing if the effect the parents education level has on their childs performance changes as the child gets older. It would therefore make sense to stratify on which level the school is, elementary school, middle school or high school. We can then see if the estimated slope coefficient is different in each of these strata. The sampling frame has* 4421 *elementary schools,* 755 *middle schools and* 1018 *high schools and we choose to sample* 50 *schools from each strata.*

Table 3 shows the estimated slope with 95% confidence intervals along with true slopes for each school level. We observe that all the true values are inside the confidence intervals. We can see that parent education level seems to have a higher effect in Elementary schools than higher school levels. Regression on all the strata together gives the slope 158*, which is somewhere in between all the individual slopes. Not doing stratification would make us loose the information for each school type.*

4.2 Clustering

In clustering we split the sampling frame into a partition as in stratification. Here, however, we do a probability sample to choose which of the subsets we will collect data from. Then we have to do a probability sample inside of each of these chosen subsets, $S_1, S_2, \ldots, S_n$.

Definition 5. ***Primary sampling unit (psu)*** *are subsets of the sampling frame, that are sampled first in a sampling scheme. These are also often called* ***clusters****.*

There are two types of cluster sampling; one-stage cluster sampling and multi-stage cluster sampling. In one-stage cluster sampling we sample all the elements in the chosen subsets, or clusters, of S, so each element in these clusters have probability 1 of being included. In multi-stage cluster sampling, however, we make a sample of the **secondary sampling units (ssus)**, which are individuals inside the clusters. Here not all ssus, or individuals, inside the clusters are included. Multi-stage cluster sampling is often what is used in practice, but since the ideas are similar to one-stage cluster sampling and the formulas get much more complicated in multi-stage cluster sampling, we will restrict outselves to one-stage cluster sampling in this thesis.

Let N be the number of clusters (psus) in the population and let $M_i,\ i = 1, 2, \ldots, N$ be the number of individuals (ssus) in each cluster. Let $t_{y,i}$ be the total of the y's in cluster i. Since we sample all the ssus inside the clusters, we know the totals for the clusters in the sample. We can therefore consider each cluster a sampling unit as we have done earlier in the thesis. Instead of letting the y's be the sampled values, we let $t_{y,i}, i \in S$ be the sampled values. In one-stage cluster sampling, an SRS is usually taken to choose which clusters to include in the sample. This gives a similar estimator for the total as for a SRS, except that the sampling units here are the totals for each cluster instead of individuals as in a normal SRS, $\hat{t}_y = \sum_{i \in S} \frac{N}{n} t_{y,i} = \sum_{i \in S} w_i t_{y,i}$. This also gives the same estimate for the variance as in an SRS, $\widehat{\text{Var}}\left(t_y\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) \hat{\text{Var}}\left(t_{y,i}\right) = \frac{N^2}{(n-1)n}\left(1 - \frac{n}{N}\right) \sum_{i \in S} \left(t_{y,i} - \hat{\bar{t}_y}\right)^2$, where $\hat{\bar{t}_y} = \frac{1}{n} \sum_{i \in S} t_{y,i}$.

Larger populations often mean that the totals for that population is also large, for example, when we are measuring API scores for schools, the total of all the schools in a population will increase as the number of schools we sample increases, as the score is always positive. This means that the variance of the total estimates will usually increase if the clusters have very different population sizes. It can therefore be useful to try to keep the population sizes of the different clusters as similar as possible. But one should be careful not to put too much weight on keeping the population sizes equal, as that may cause some clusters to become too large physically which will make it expensive to measure all individuals in it, which defeats the purpose of clustering in the first place.

A cluster sample will almost always have a larger variance than a SRS of the same size. This is because individuals inside a cluster will usually be more similar than individuals across clusters. This means that measuring say 10 people inside one cluster will give less information than measuring 10 people randomly chosen in the whole population. This causes cluster sampling to have a larger uncertainty. Measuring inside a cluster is often cheaper than measuring people sampled randomly from the whole population, though. For example, if the clusters are geographical areas, which they often are, and each individual sampled needs a researcher or interviewer to physically travel to them, then one will save time and money not having to travel to so many different places. This means that it is often possible to sample many more individuals in cluster sampling than when doing samples from the whole population, which will decrease the uncertainty and make cluster sampling more competitive with other sampling methods.

Example 4. *We again consider the API dataset and want to see if average class size for kindergarden through third grade impacts the API score for a school. Suppose that to collect the data, an interviewer has to travel to each school. Then it would take much time and become expensive if the interviewer had to travel to random schools all across California. It would be cheaper and take less time to collect data in just some school districts. We therefore choose to cluster on the school districts. There are* 757 *school districts in California and we take a SRS to get a sample of size* 15*. The resulting sample of clusters gives us a sample of* 183 *schools. Based on the data collected from these schools we can now make a regression line and confidence intervals for the regression coefficients. To make the regression line and to accurately estimate the variance we have to take the correlation between the individuals in the same cluster into account. If we do not take this into account we will get a different result and a too small variance estimate.*

Observe from Table 2 that the standard error of the model where the clustering is ignored is significantly smaller than the standard error of the model where the clustering is accounted for. This is because we assume schools inside each district are more similar than schools in different districts. This causes a sample with many schools in the same districts to carry less information about the whole population than one where schools in all districts have a chance of being included in the sample. The different standard errors means that the hypothesis test, where the null-hypothesis is that the slope is zero, gives different results for the two models. Ignoring

	Slope	Std. Error	P-value
Clustering taken into account	3.232	9.254	0.732
Clustering not taken into account	3.232	0.593	$7.49 * 10^{-8}$

Table 2: Slope for regression lines along with uncertainty and P-value for the null-hypothesis that the slope is zero. The first row is for the case where the clustering is taken into account. In the second row the clustering was ignored, the samples were assumed to be independent.

the clustering leads to the belief that larger average class size for kindergarden through third grade leads to a higher API score for the school. The sample does, however, suggests no such thing, as the P-value of the model where the clustering is taken into account is 0.732.

4.3 Complex surveys

In practice, we often combine clustering and stratification to what is called a **complex survey**. We usually first create strata for the different subpopulations we are interested in data from. Then we use clustering to make it practically possible to perform the survey.

A real example, where survey statistics with a complex survey design is used, is the Demographic and Health Survey (DHS) in Kenya. This survey collects, among other things, information regarding the birth rates and mortality rates in Kenya. This type of survey is important because no one has any overview of vital statistics in many developing countries. Ideally, one would of course do a full census of the population, but that is not possible in practice as it would be extremely expensive. Each household that is surveyed has to be visited by a trained interviewer. Visiting households that are spread through the whole country would need both much travel time and many interviewers who have to be trained. Instead one designs a complex survey combining the sampling methods introduced earlier in Section 3. This is cheaper, and if the survey is well designed one will get good data as well.

In the DHS in Kenya the researchers wanted data on both urban and rural populations in each county, so they first started with splitting the country into strata. Two strata for each county, one for the urban population and one for the rural population, except for two counties which only have an urban population. Each of these strata were then further divided into smaller geographic units, these are the clusters. The researchers then do a SRS of the clusters inside each stratum to choose which ones to visit. Kenya National Bureau of Statistics et al. ((2015))

Inside each cluster the researchers then sample 25 households to actually visit. The fact that the researchers make a new sample inside the cluster is, as mentioned in Section 4.2, called two-stage clustering. It has not been a topic of this thesis, but the principles used are the same as the ones used in one-stage clustering. Kenya National Bureau of Statistics et al. ((2015))

Counterintuively, one can often get better results doing a smaller well designed survey than sampling as many as possible. This is because, if we have a small sample, we need fewer interviewers and can therefore make sure they are all well trained and suited to the job. If we just focus on getting as large samples as possible one often has to compromise on the quality of the training the interviewer get. This can lead to poorly trained interviewers introducing errors to the data. For example, if no one from the household they are supposed to visit are home, they might visit their neighbours instead of coming back later. This compromises the results of the survey, as some demographics might be more likely to be home than others.

5 Variance estimation

A central part of doing statistics is the calculation of uncertainties. We usually express the uncertainty by the variance. Without an idea of the variance of an estimated quantity, the estimate is all but useless, as the variance could be so large that almost any confidence interval will include all possible values. It is therefore important to have a reasonable estimate of the variance. The problem, however, is that in regression the values we are interested in are non-linear expressions of observations. We therefore can not find a closed form expression for the variance estimate using normal rules, so we instead have to use other techniques to find variance estimates. One of the more commonly used methods in survey statistics, and one which will give a closed form expression for the estimate, is linearization.

Example 5. *Let* $\widehat{W} = \frac{\hat{t}_y}{\hat{t}_x}$*, where* $\hat{t}_y$ *and* $\hat{t}_x$ *are stochastic variables. We are interested in finding* $\operatorname{Var}\left(\widehat{W}\right) = \operatorname{Var}\left(\frac{\hat{t}_y}{\hat{t}_x}\right)$*. Since this is a non-linear expression, there is no easy formula to calculate it. We do, however, know how to calculate the variance of a linear combination. We can therefore take advantage of the fact that we can approximate* $\widehat{W} = \frac{\hat{t}_y}{\hat{t}_x}$ *by using a first degree Taylor approximation, which is linear.*

Let $h(a,b) = \frac{a}{b}$*. Then* $\widehat{W} = h(\hat{t}_y, \hat{t}_x)$ *and* $h'(a,b) = \begin{bmatrix} \frac{1}{b} & -\frac{a}{b^2} \end{bmatrix}$*.*

This means that

$$\begin{aligned}\widehat{W} = h(\hat{t}_y, \hat{t}_x) &\approx h(t_y, t_x) + h'(t_y, t_x)\left(\begin{bmatrix}\hat{t}_y \\ \hat{t}_x\end{bmatrix} - \begin{bmatrix} t_y \\ t_x\end{bmatrix}\right) = \frac{t_y}{t_x} + \begin{bmatrix}\frac{1}{t_x} & -\frac{t_y}{t_x^2}\end{bmatrix}\begin{bmatrix}\hat{t}_y - t_y \\ \hat{t}_x - t_x\end{bmatrix} \\ &= \frac{t_y}{t_x} + \frac{1}{t_x}\left(\hat{t}_y - t_y\right) - \frac{t_y}{t_x^2}\left(\hat{t}_x - t_x\right)\end{aligned}$$

Since t_y *and* t_x *are deterministic values, they have zero variance and we can treat them as any other constants in the variance expression. We can therefore apply the standard rules for calculating variances of linear expressions and we get that*

$$\operatorname{Var}\left(\widehat{W}\right) \approx \operatorname{Var}\left(\frac{t_y}{t_x} + \frac{1}{t_x}\left(\hat{t}_y - t_y\right) - \frac{t_y}{t_x^2}\left(\hat{t}_x - t_x\right)\right) = \frac{1}{t_x^2}\operatorname{Var}\left(\hat{t}_y\right) + \frac{t_y^2}{t_x^4}\operatorname{Var}\left(\hat{t}_x\right) - 2\frac{t_y}{t_x^3}\operatorname{Cov}(\hat{t}_x, \hat{t}_y),$$

which gives the simple expression

$$\widehat{\operatorname{Var}}\left(\widehat{W}\right) = \frac{1}{\hat{t}_x^2}\widehat{\operatorname{Var}}\left(\hat{t}_y\right) + \frac{\hat{t}_y^2}{\hat{t}_x^4}\widehat{\operatorname{Var}}\left(\hat{t}_x\right) - 2\frac{\hat{t}_y}{\hat{t}_x^3}\widehat{\operatorname{Cov}}(\hat{t}_x, \hat{t}_y)$$

Taylor approximation works best if the point we are approximating around is near the point we want the function value at. This means that in this case we should choose a fixed point near $(\hat{t}_x, \hat{t}_y)$*. Since this is an approximation of* (t_x, t_y)*, that point is a natural choice to approximate around. Another advantage of approximating around* (t_x, t_y) *is the fact that to get a variance approximation, we can exchange* t_x *and* t_y *with* $\hat{t}_x$ *and* $\hat{t}_y$ *and get a numeric value. Since* $\hat{t}_y$ *and* $\hat{t}_x$ *are linear expressions we are able to find closed form expressions for their variance estimates and we can therefore calculate* $\widehat{\operatorname{Var}}\left(\widehat{W}\right)$*.*

To illustrate this with a concrete example, suppose $\hat{t}_x = 100$ *with* $\widehat{\operatorname{Var}}\left(\hat{t}_x\right) = 5$ *and* $\hat{t}_y = 230$ *with* $\widehat{\operatorname{Var}}\left(\hat{t}_y\right) = 23$ *and that they are independent such that* $\operatorname{Cov}\left(\hat{t}_x, \hat{t}_y\right) = 0$*.*

Then we have $\widehat{\operatorname{Var}}\left(\widehat{W}\right) = \frac{1}{100^2}23 + \frac{230^2}{100^4}5 = 0.0049$*.*

More generally, let $h(x)$, where $x = (x_1, x_2, \ldots, x_p)$, be a function such that $h(x) = \theta$, where θ is the quantity we are interested in. Further, let $\hat{x}$ be an estimate of x, such that $h(\hat{x}) = \hat{\theta}$. If h is linear, there is no problem finding a closed form for the variance estimate, so we assume here that h is a non-linear function.

By Taylor's theorem, we know that $h(\hat{x}) = h(x) + h'(x)(\hat{x} - x) + \int_x^{\hat{x}} (\hat{x} - t)h''(t)dt$, Lindstrøm ((2016)), where the last term is usually small compared to the first two in statistics, ((Lohr, 2009, Chapter 9)). This allows us to use the approximation $h(\hat{x}) \approx h(x) + h'(x)(\hat{x} - x)$. $h(x) = \left[\frac{\partial h}{\partial x_1}(x_1) \quad \frac{\partial h}{\partial x_2}(x_2) \quad \ldots \quad \frac{\partial h}{\partial x_p}(x_p)\right]$, which means that $h(\hat{x}) \approx h(x) + h'(x)(\hat{x} - x) = h(x) + \sum_{i=1}^{p} \frac{\partial h}{\partial x_i}(x_i)\,(\hat{x}_i - x_i)$, which is a linear expression in $\hat{x}$. Using known rules for variance estimation of linear expressions we have

$$\begin{aligned}
\text{Var}\left(\hat{\theta}\right) = \text{Var}\left(h(\hat{x})\right) &\approx \text{Var}\left(h(x) + \sum_{i=1}^{p} \frac{\partial h}{\partial x_i}(x_i)\,(\hat{x}_i - x_i)\right) \\
&= \sum_{i=1}^{p} \left(\frac{\partial h}{\partial x_i}(x_i)\right)^2 \text{Var}\left(\hat{x}_i\right) + \sum_{i \neq j} \frac{\partial h}{\partial x_i}(x_i)\frac{\partial h}{\partial x_j}(x_j)\text{Cov}\left(\hat{x}_i, \hat{x}_j\right)
\end{aligned}$$

We do not, however, know x, as that is the true values we are trying to estimate. But by replacing x with $\hat{x}$, and by replacing the variances of $\hat{x}$ with estimated variances in the expression we get an estimate of the variance

$$\widehat{\text{Var}}\left(\hat{\theta}\right) = \sum_{i=1}^{p} \left(\frac{\partial h}{\partial x_i}(\hat{x}_i)\right)^2 \widehat{\text{Var}}\left(\hat{x}_i\right) + \sum_{i \neq j} \frac{\partial h}{\partial x_i}(\hat{x}_i)\frac{\partial h}{\partial x_j}(\hat{x}_j)\widehat{\text{Cov}}\left(\hat{x}_i, \hat{x}_j\right)$$

In the regression case, the differentiation and the estimates of covariance are complicated, so we will not show the derivation here. However, it can be shown that

$$\begin{aligned}
\text{Var}(\hat{B}_1) &= \text{Var}\left(h(\hat{t_{xy}}, \hat{t_x}, \hat{t}_y, \hat{t}_{x^2}, \hat{N}))\right) \\
&\approx \frac{\widehat{\text{Var}}\left(\sum_{i \in S} w_i q_i\right)}{\left(\sum_{i \in S} w_i x_i^2 - \frac{\left(\sum_{i \in S} w_i x_i\right)^2}{\sum_{i \in S} w_i}\right)^2}
\end{aligned}$$

where $q_i = (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}})$, and $\bar{x} = \frac{\hat{t}_x}{\hat{N}}$, see ((Lohr, 2009, Chapter 11)) for more details.

6 Discussion

When using sampling designs where the sampling probabilities differ, we risk the chance of having a bias in our results if we do not take the sampling design into account. This is illustrated in Example 1 where the regression line where we ignore the unequal sampling probabilities is too flat, and in Example 2, where we get different estimates of the totals. A bias in the regression line can quickly make one reach conclusions, about the relationship between the predictor and the response, not actually supported by the data.

Since we can use weights, calculating estimates of most quantities is straight forward, including estimates of regression coefficients. The non-linear nature of the regression coefficients, however, makes it difficult to estimate the variances. There are several methods to get variance estimates, including using resampling techniques, but the one used in this thesis is linearization. Using linearization allows us to get reasonable variance estimates for the regression coefficients.

As seen in Section 5 linearization has given us a closed form expression for the variance estimates of the slope parameter, and a similar one for the intercept is also possible to derive.

Ignoring sampling design can cause major problems with variance estimation. In many sampling designs, clustering is used, which means the sampled units are correlated. This means that the sample carries less information than if the units were sampled independently. Assuming that the sampled units are independent when they are not will cause underestimation of the variance. Believing the variance is smaller than it actually is can lead to a stronger belief in the results than warranted. In regression one is often interested in whether the regression line is flat or not, as a flat regression line means no relationship between predictor and response. Having a too small variance estimate might cause us to reject the null hypothesis that the slope is zero, while we with a more correct variance would not reject that null hypothesis. Ignoring the fact that one samples from a finite population might cause the opposite problem: believing that one has more uncertainty than there actually is. This happens when one forgets to take the fpc into account, which reduces the variance as one samples a larger proportion of the population. This would reduce the power of the hypothesis test as it would become harder to reject the null hypothesis, even when it is false.

It is important to note that using survey statistics will never give wrong results. As illustrated in the bottom left panel of Figure 2, the regression line from classical linear regression is the exact same as the one from survey statistics where we assume each observation is independently sampled. This is also true in general. A disadvantage of survey statistics, however, is that one in general needs more data to achieve good results. This is because in model based statistics, the assumptions one make gives additional structure to the data. This is structure which we would have to "estimate" using survey statistics, causing us to "use up" some of the statistical power we want to use for estimating regression coefficients and variance. If we do not have enough data to get robust conclusions we would need to collect more data if we want to keep using survey statistics. Another option, however, if it looks like the assumptions regarding the distribution is fulfilled, is to switch to a model based regression. In addition, one would have to make sure that clustering and stratification will not cause problems by violating independence. If observations inside clusters are no more similar than observations in different clusters one can disregard the clustering and still get good results. One can correct for different sampling probabilities by incorporating weights in the model based regression. The correct choice is always to use survey statistics, but if you don't have enough data and the model seems to fit it can sometimes be an acceptable alternative to use model based regression instead.

References

Kenya National Bureau of Statistics, Ministry of Health, National AIDS Control Council, Kenya Medical Research Institute, National Council for Population and Development, The DHS Program, and ICF International. Kenya demographic health survey 2014, 2015.

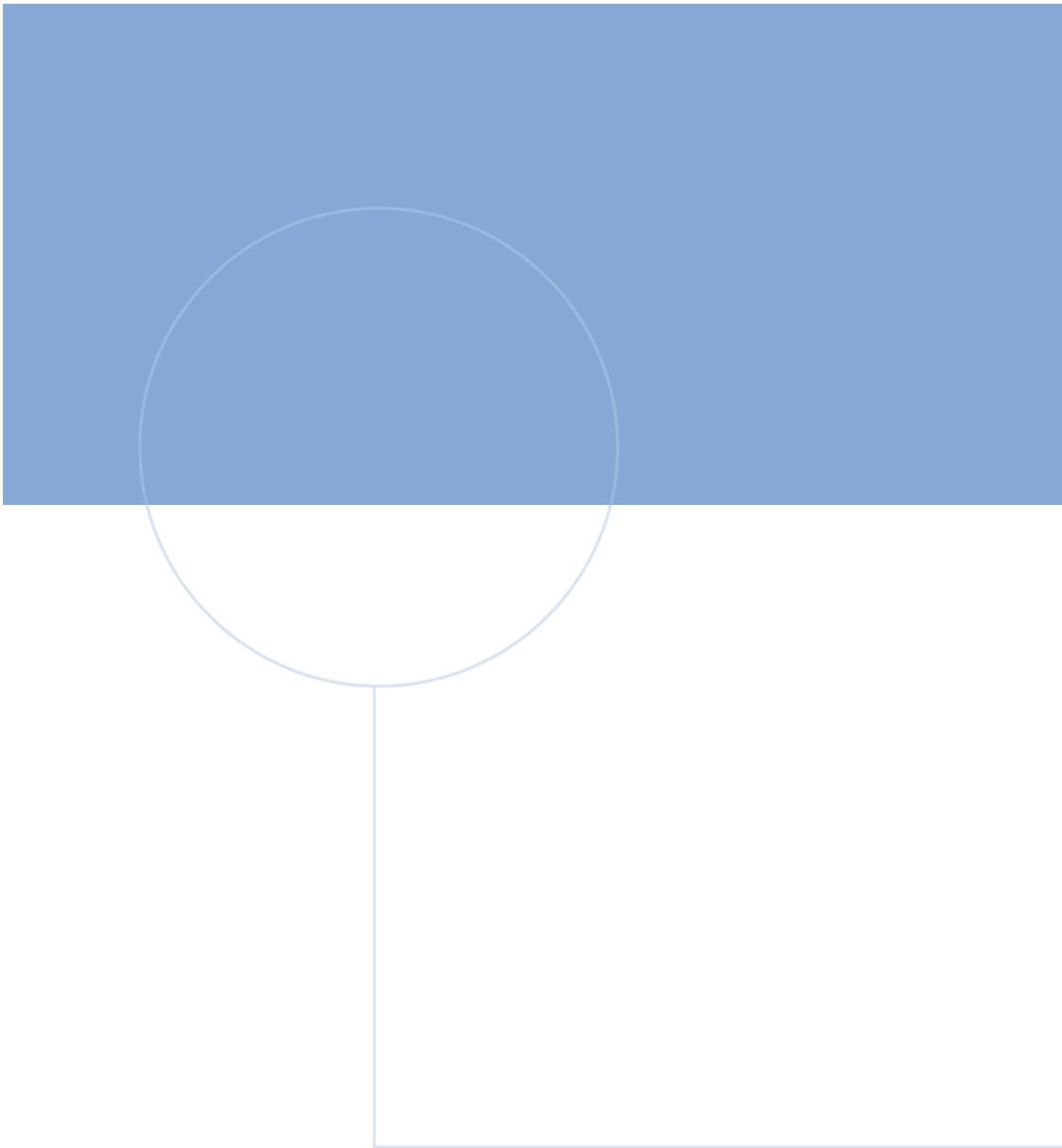
R. J. Larsen and M. L. Marx. *An Introduction to Mathematical Statistics and it's Applications.* Fifth edition, 2012.

T. L. Lindstrøm. *Kalkulus.* Universitetsforlaget, Oslo, Norway, fourth edition, 2016.

S. L. Lohr. *Sampling: Design and Analysis.* Brooks/Cole, Boston, Massachusetts, second edition, 2009.

T. Lumley. *Complex Surveys, A Guide to Analysis Using R.* John Wiles & Sons, Hoboken, New Jersey, 2010.

T. Lumley. survey: analysis of complex survey samples, 2020. R package version 4.0.



NTNU

Norwegian University of
Science and Technology