Imputation for Nonresponse using the Annual Financial Statistics Survey

By

Smeeta Singh

Imputation for Nonresponse using the Annual Financial Statistics Survey

By

Smeeta Singh

Submitted in fulfilment of the requirements for the degree of Master of Science, in the School of Statistics and Actuarial Science at the University of KwaZulu-Natal.

June 2011

Abstract

In this dissertation, we focus on the Annual Financial Statistics (AFS) survey. This is a survey conducted by Statistics South Africa, the national statistics office of South Africa. The main purpose of this survey is to provide information for compiling of the GDP estimates, value-added and its components, which are used to monitor and develop government policies. The AFS covers a sample of private enterprises operating the formal non-agricultural business sector of the South African economy, excluding financial, insurance and government institutions.

Quality is a key area of importance in this organisation and therefore methodology and standards need to be monitored, evaluated and reviewed on a regular basis. This would assist in ensuring that Statistics South Africa is following international best practices for collection and estimation of official statistics. We focus on nonresponse for the Annual Financial Statistics survey, and investigate an alternative method for adjusting for nonresponse and in particular focus on improving the method of dealing with nonresponse thereby improving the estimates from the AFS survey.

Declaration

This dissertation represents original work by the author and has not otherwise been submitted in any form for any degree or diploma to any tertiary institution. Where the use has been made of the work of others, it is duly acknowledged in the text.

S Singh

Acknowledgements

I would like to thank the following people each of whom has made this work possible:

Firstly, I would like to thank my supervisor, Prof D North, for her time and expert guidance as well as my co-supervisor, Prof T Zewotir for his input.

My sincerest thanks go to my parents and my sisters for their support, understanding and encouragement.

Thanks to Statistics South Africa for creating the opportunity for me to pursue my post graduate studies.

CONTENTS

Introduction

Nonresponse is a pervasive fact of life in any survey area. Elliot (1991) explains that by adopting a standard and well researched set of field procedures and question design methods, much can be done to minimise it and its effects. However, beyond a certain point the additional effort entailed in pursuing the last non-respondent is simply not cost effective. An alternate strategy is to adopt procedures that aim to correct for the effects of the residual nonresponse during the analysis stage of the survey. One such strategy is to fill in or ñimputeò missing values. There is a variety of imputation methods, ranging from extremely simple to rather complex methods. Imputation methods preserve the full sample size, which can be advantageous for both precision and biasness.

Official statistics is the life-blood of any countryôs economy as well as society. Statistics South Africa (Stats SA) is the national statistics office for South Africa. Each year, Stats SA produces more than 250 statistical releases covering a broad spectrum of information on the South African society and the economy. Every day social and economic decisions are made within government, the private sector and the broader community based on official statistics. The Annual Financial Statistics (AFS) survey is one of the economic surveys carried out by Stats SA. The AFS is designed to give financial information for the private enterprises operating in the formal non-agricultural business sector of the South African economy, excluding financial intermediation, insurance and government institutions.

In this dissertation we focus on the methods used to deal with nonresponse for the AFS survey. We discuss the current methods used for adjusting for missing values (nonresponse) in the AFS, followed by alternative methods for adjusting for nonresponse. We investigate the effect of this current method of correcting for nonresponse in the AFS and will compare that with other ways of correcting for missing values.

In chapter 1, we get a better understanding of Stats SA, the national statistics office, as well as the AFS survey. Chapter 2 covers the principle steps in a sample survey, and concentrates on simple random sampling without replacement and sampling using stratification as the AFS survey uses these sampling methods.

Chapter 3 defines nonresponse and explains procedures to deal with nonresponse including estimation in the presence of nonresponse. Focus will be on historical imputation and ratio imputation as these imputation methods are used for the AFS survey. In chapter 4 we investigate the regression imputation method as an alternative imputation technique for the correction for missing data in the AFS survey.

Chapter 5 entails the application in respect of the different approach to handling nonresponse and comparison of the two methods. The conclusions look at steps forward with regard to the new methods being investigated as well as the probability of implementation of these techniques at Statistics South Africa.

Chapter 1

The Annual Financial Statistics survey

The Annual Financial Statistics (AFS) survey is an annual survey conducted by Statistics South Africa (Stats SA), the national statistics office of South Africa. This chapter gives one a better understanding of Stats SA as well as the AFS survey.

Stats SA is a government organisation under the National Planning Commission. The strategic direction of Stats SA may be found in Statistics South Africa (2008)[B], and is informed by its vision, which is to be ñYour leading partner in quality statisticsò, providing stakeholders and the public with high quality statistical information. The mission of Stats SA is ñTo lead and partner in statistical production systems for evidence-based decisionsò. The Annual Financial Statistics survey falls in the Private Sector component under the Financial Statistics division in the Economic Statistics cluster and is widely used by government, analysts and researchers.

Quality is an essential element in all areas of practise of a national statistics office. It is thus important that measures and processes are in place to ensure that the highest quality is attained. In its endeavour to fulfil the purpose of providing users with quality information, Stats SA has adopted the following principles developed by the Economic and Social Council Statistical Commission of the United Nations http://www.statssa.gov.za/ .

- Relevance, impartiality and equal access;
- Professional standards and ethics;
- Accountability and transparency;
- Prevention of misuse;
- Cost-effectiveness;
- Confidentiality;
- Legislation;
- National coordination;
- International standards; and

- International cooperation.

1.1 What is the Annual Financial Statistics survey?

In 1998, the Australian Bureau of Statistics (ABS) assisted Stats SA to examine the economic statistics strategy in terms of key economic indicator collections. The annual Economic Activity Survey (EAS) was subsequently introduced in 1998 as it was considered to be an important survey to collect annual financial information of businesses in the economy. In 2007, the EAS was renamed as the Annual Financial Statistics (AFS) survey, for more clarity and understanding of the survey. The AFS survey follows international best practices with regard to the methodology and the implementation of the survey.

The AFS covers a sample of private enterprises operating in the formal non-agricultural business sector of the South African economy, excluding financial intermediation, insurance and government institutions. The AFS is designed to give information on income and expenditure items, as well as capital expenditure on new and existing assets, balance sheet items and the carrying value of property, plant and equipment and intangible assets at the end of the financial year for the South African enterprises selected to be part of the survey (selection process to be discussed).

The main purpose of the survey is to provide relevant information for government to make informed economic policy decisions (Statistics South Africa, 2007). The AFS is used to ï

- compile estimates of the GDP, value-added and its components, which are used to monitor and develop government policy;
- derive a set of economic measures based on information available from the standard financial accounts of trading and employing businesses;
- provide data which feed into the Supply and Use tables (SU-tables) and the annual National Accounts;
- provide economic indicator statistics;

- provide the private sector with the necessary information in the analysis of comparative business and industry performance;
- provide a convenient vehicle to collect more detailed and specific information regarding economic data; and
- be instrumental in reducing problems with lags in national accounts data.

The statistical unit for the collection of information for the AFS will be referred to as an enterprise. An enterprise is thus defined as a legal unit or a combination of legal units that includes and directly controls all functions necessary to carry out its production activities. Each enterprise is classified to an industry that reflects the main activity of the enterprise, for example, an enterprise could have activities in both manufacturing and retail trade, in which case this enterprise would be classified to the industry where its activity is more predominantly focused.

Stats SAôs classifications are based on the Standard Industrial Classification for all Economic Activities (SIC) which is based upon the United Nationôs International Standard Industrial Classification of all Economic Activities (ISIC) (Statistics South Africa (2006)).

The SIC is a standard classification of productive, economic activities of industries. The objective of the SIC is to classify collected statistical information as far as possible according to categories of activities which are as homogeneous (of the same kind) as possible. In Statistics South Africa (2006), an industry is defined as the set of all production units engaged primarily in the same or similar kinds of productive economic activities. The term óindustryô is used in the widest sense to cover all economic activities from the primary industries (agriculture, forestry, fishing and mining) to secondary industries (manufacturing, electricity, gas and water supply and construction) and includes the rendering of tertiary services (wholesale, retail and motor trade and repairs for it, including personal and household goods, hotels and restaurants, transport, storage and communication, financial intermediation, insurance, real estate and other business services, community, social, and personal, recreational, cultural services and other private households, etc. and other activities not adequately defined).

According to SIC, enterprises in South Africa are grouped into nine industries on a 1-digit level.

Table 1.1: The 1-digit (major division) classification of industries according to the Standard Industrial Classification for all Economic Activities (SIC)

SIC	Industry
1	Forestry and fishing
2	Mining and quarrying
3	Manufacturing
4	Electricity, gas and water supply
5	Construction
6	Trade
7	Transport, storage and communication
8	Real estate, activities auxiliary to financial intermediation and other business services, excluding financial intermediation and insurance
9	Community, social and personal services, excluding government institutions

Enterprises are classified in more detail at a lower digit level. Identification of economic activities for enterprises is as follows:

1. **Major division** (1-digit) of SIC.
2. **Division** (2-digit) of SIC.
3. **Major group** (3-digit) of SIC.
4. **Group** (4-digit) of SIC.
5. **Subgroup** (5-digit) of SIC.

Table 1.2: Examples of Standard Industrial Classification of all Economic Activities (SIC)

Title of category	Major division	Division	Major group	Group	Sub group
MANUFACTURING	3				
MANUFACTURE OF FOOD PRODUCTS, BEVERAGES AND TOBACCO PRODUCTS		30			
PRODUCTION, PROCESSING AND PRESERVATION OF MEAT, FISH, FRUIT, VEGETABLES, OILS AND FATS			301		
Production, processing and preserving of meat and meat products				3011	
Slaughtering, dressing and packing of livestock, including poultry and small game meat					30111
Manufacture of prepared and preserved meat, including sausage; by-products (hides, bones, etc.)					30112
Production of lard and other edible fats					30113
Processing and preserving of fish and fish products				3012	
Manufacture of canned, preserved and processed fish, crustacean and similar foods (except soups)					30120
TRANSPORT, STORAGE AND COMMUNICATION	7				
LAND TRANSPORT: TRANSPORT VIA PIPELINES		71			
RAILWAY TRANSPORT			711		
Rail transport				7111	
Inter-urban railway transport					71111
Railway commuter services					71112
OTHER LAND TRANSPORT			712		
Other scheduled passenger land transport (except cable, funicular)				7121	
Urban, suburban and inter-urban bus and coach passenger lines					71211
School buses and other scheduled transport, except cable, funicular					71212

Table 1 shows examples of activities classified under the Standard Industrial Classification of all Economic Activities (SIC).

Four group sizes are specified per industry; namely size group i, $i \in \{1, 2, 3, 4\}$; using the cut-off points from the Department of Trade and Industry (DTI) (see table 1.3). Each enterprise would fall into a group size per industry type according to their measure of size. For example, if a construction enterprise had a measure of size between R9 million and R39 million, then this enterprise would fall in size group 2.

Table 1.3: Cut-off points for the various size groups by industry Source: Trade and Industry (2003), National Small Business Amendment Bill - DTI 2003 (Factor adjusted by Stats SA for AFS 2008)

Industry	AFS - Enterprise size			
	Large	Medium	Small	Very small
	Size group 1	Size group 2	Size group 3	Size group 4
	Rand			
Forestry and fishing	Turnover > 7 500 000	Turnover > 4 500 000 , Turnover Ò 7 500 000	Turnover > 750 000 , Turnover Ò 4 500 000	Turnover Ò 750 000
Mining and quarrying	Turnover > 58 500 000	Turnover > 15 000 000 , Turnover Ò 58 500 000	Turnover > 6 000 000 , Turnover Ò 15 000 000	Turnover Ò 6 000 000
Manufacturing	Turnover > 51 000 000	Turnover > 13 000 000 , Turnover Ò 51 000 000	Turnover > 5 000 000 , Turnover Ò 13 000 000	Turnover Ò 5 000 000
Electricity, gas and water supply	Turnover > 76 500 000	Turnover > 19 500 000 , Turnover Ò 76 500 000	Turnover > 7 650 000 , Turnover Ò 19 500 000	Turnover Ò 7 650 000
Construction	Turnover > 39 000 000	Turnover > 9 000 000 , Turnover Ò 39 000 000	Turnover > 4 500 000 , Turnover Ò 9 000 000	Turnover Ò 4 500 000
Wholesale trade	Turnover > 96 000 000	Turnover > 48 000 000, Turnover Ò 96 000 000	Turnover > 9 000 000 , Turnover Ò 48 000 000	Turnover Ò 9 000 000
Retail and Motor trade	Turnover > 58 500 000	Turnover > 28 500 000 , Turnover Ò 58 500 000	Turnover > 6 000 000 , Turnover Ò 28 500 000	Turnover Ò 6 000 000
Accommodation and catering	Turnover > 19 500 000	Turnover > 9 000 000, Turnover Ò 19 500 000	Turnover > 7 650 000, Turnover Ò 9 000 000	Turnover Ò 7 650 000
Transport, Storage and communication	Turnover > 39 000 000	Turnover > 19 500 000 , Turnover Ò 39 000 000	Turnover > 4 500 000, Turnover Ò 19 500 000	Turnover Ò 4 500 000
Real estate, activities auxiliary to financial intermediation and other business services (excluding financial intermediation & insurance)	Turnover > 39 000 000	Turnover > 19 500 000 , Turnover Ò 39 000 000	Turnover > 4 500 000, Turnover Ò 19 500 000	Turnover Ò 4 500 000
Community, social and personal services (excluding government institutions)	Turnover > 13 000 000	Turnover > 6 000 000 , Turnover Ò 13 000 000	Turnover > 1 000 000, Turnover Ò 6 000 000	Turnover Ò 1 000 000

Industry and Trade is another division under the Economic Statistics cluster of Stats SA. Under this division Large Sample Surveys (LSS) are carried out. The LSS focus on specific industries each year and collect product information in addition to the financial information collected by the AFS. In 2006 it was decided that the two divisions, AFS and LSS, should work together to reduce respondent burden. This resulted in financial information on industries surveyed by LSS and not AFS, being included in the AFS publication. In this dissertation we will look at data and estimates for the AFS and LSS for 2008. In 2008 the LSS division collected data for the manufacturing industry and community, social and personal services, excluding government institutions industry while the remaining industries were collected by AFS.

Estimates calculated from the AFS survey is published at industry level (see table 1.1). Disaggregated industry estimates, estimates per business size and estimates per organisational type are also available on the Stats SA website (http://www.statssa.gov.za/) and are up-dated on a yearly basis. Disaggregated industry estimates are estimates broken down into more detailed classifications e.g. the mining and quarrying estimates are broken down, showing the estimates for the different types of mining and quarrying such as mining of gold and uranium ore, mining of platinum group metals, mining of copper, quarrying of limestone and limeworks etc. Estimates per business size, group the estimates according to the different size groups. Estimates per organisational types, group the estimates according to the different organisational types e.g. private companies, close corporations, public companies etc.

Since many enterprises revise their data, revisions are made annually by Stats SA to the published data. This however, is done with extreme caution and monitored closely. The estimates that are published are thus the preliminary estimates and are released along with the revised estimates for the year prior. For example, when the AFS 2008 results were published, these were the preliminary estimates for 2008 and the estimates for 2007 were revised and published in the same publication.

1.2 Sampling for the Annual Financial Statistics survey

The Methodology and Evaluation division, under the Methodology and Standards cluster, draws the sample for the AFS survey. The sampling specifications for the survey are determined by National Accounts division under the Economic Statistics cluster. The survey area proposes a sample size and sampling precision level.

The Business Register division receives the snapshot from the South African Revenue Services (SARS). A snapshot is a list of all tax paying businesses in South Africa at a specific point in time. Statistics South Africa (2008)^{A}, explains how a common frame is created by removing ceased, deactivated and duplicated enterprises. From the common frame, the financial frame is created by removing the enterprises which are only income tax sourced. Percentile cut-offs are created for each industry and enterprises below the percentile cut-off are removed from the frame. This is done to avoid sampling very small enterprises. These enterprises are accounted for by introducing a factor which is multiplied to the weights for size group 4 enterprises. The business sampling frame is further adjusted by removing enterprises not classified according to the specified SIC digit-levels according to the sampling specifications.

The AFS survey is sampled using a stratified random sample design. This is based on business turnover from the business sampling frame as a measure of size. Each enterprise would fall into a group size per industry type according to their measure of size. All enterprises in the group size representing the large enterprises (size group one) are selected in the sample. This is due to the size group one enterprises having a large impact on the economy. For AFS 2008, the size group one enterprises contributed 80% of the total turnover for all industries. For size groups two, three and four, enterprises on the business sampling frame are placed into strata, from which a random sample is selected. Statistics South Africa (2009), shows, that each stratum contains two parts of information, namely the industrial sector (primary stratum) to which an enterprise belongs and the group size (secondary stratum) to which an enterprise is assigned. The industrial sector is derived from the SIC variable and the group size based on the measure of size (turnover). The sampling frame thus consists of all enterprises classified in strata according to their industrial sector then a secondary classification of strata according to enterprise size. The Neyman optimal allocation is used to select samples from each stratum.

Samples are chosen once a year. Since samples are drawn for size group two, three and four, weights are applied to results to adjust for all the enterprises in the population. Weights can be large due to sample design and nonresponse. For example, if the size group 2 population for the forestry and fishing industry had 123 enterprises and 10 were sampled, then the design weight would be 12,3 (123 / 10). If from the 10 enterprises sampled, 6 respond then the new weight would be 20,5 [(123 / 10) * (10 / 6)] (see chapter 3 ï Missing data). Table 1.5 shows the percentage per size group that needed to be compensated for. From the table we can see that for AFS 2008, weights were adjusted to compensate for 26% of the size group 2 sample.

Table 1.4 shows the sampling specifications for some of the industries for the AFS 2008 sample. Sampling specifications are the classification level at which enterprises should be drawn for the sample per industry (see table 1.2). Higher classification levels (digit), implies greater detailed classification to be used for sampling.

Table 1.4: Example of sampling specifications for AFS 2008 sample

SIC	Industry	Levels
2	Mining and quarrying	4 digit
61	Wholesale trade	2 digit
62	Retail trade	2 digit
63	Motor trade	2 digit
64	Accommodation trade	specific 5 digit : 64101, 64102, 64103, 64109, 64201, 64202, 64203, 64204, 64209
7	Transport, storage and communication	specific 2 digit: 72, 73, 75 specific 3 digit: 711, 712, 713 specific 4 digit: (7411, 7412, 7414, 7419 grouped together), 7413

For the mining and quarrying industry, sampling was required at a 4-digit (group) SIC level. For the wholesale, retail and motor trade industry, sampling was required at a 2-digit (division) SIC level. However, for accommodation trade, sampling was required at a 5-digit (subgroup) SIC level for specific classifications. Not all the subgroups for this division were required thus the specific subgroups required are listed in the table above. In the case of transport, storage and communication, sampling was required at various levels. 2-digit SIC (division) levels were required for these specific cases: 72, 73 and 75. 3-digit (major group) SIC levels were required for these specific cases: 711, 712 and 713. At a 4-digit (group) SIC level, SIC 7413 was required. SIC 7411, 7412, 7414 and 7419 were group together for sampling. The sampling specifications are provided in accordance to the detail required in calculating the GDP.

Once the sample has been drawn, questionnaires are sent to the respective selected enterprises. Each enterprise is telephonically contacted to confirm receipt of the questionnaire. The AFS survey has a lag time of one year before the results are published, for example the AFS 2008 was released in October 2009. This is due to enterprises having different financial year ends. The advantage is that this gives the respondent sufficient time to complete their financial statements and thus complete Stats SAọs questionnaire. Various telephonic follow-ups are conducted as well as three friendly reminders are sent via email or facsimile to each of the selected enterprises to encourage responses. Visits to the respondents to assist with completion of the questionnaire are done on request. Even though several attempts are made to collect the completed questionnaire, there are still enterprises that do not respond. However, a minimum response rate of 80% has to be reached for the publication to be released.

Enterprises that were sold, closed down or liquidated during the survey period as well as enterprises that were omitted from the sample and those who have refused to complete the questionnaire contribute to the economy and are therefore compensated / imputed for. This implies that outstanding questionnaires (nonresponse) are not the only group that is compensated / imputed for. In the AFS survey, imputation is only performed on size group one enterprises which are completely enumerated, i.e. each enterprise of size group one is automatically surveyed in the AFS. For size group two, three and four, imputations are not performed for non-responding enterprises. Weights are adjusted to compensate for nonresponse for these size groups.

Table 1.5: Percentage to be compensated / imputed for AFS 2006 to 2008

AFS	**Percentage to be compensated for**			
	Size group			
	1	**2**	**3**	**4**
AFS 2006	21%	26%	29%	34%
AFS 2007	22%	23%	37%	33%
AFS 2008	22%	26%	32%	43%

Table 1.5 shows the percentage of enterprises to be compensated / imputed for AFS 2006 to AFS 2008. It can be seen that the percentage of size group one imputation is fairly consistent over the years.

Each enterprise is allocated a unit status indicator (USI) code (see table 1.6) to allow for tracking of the information, for example when a completed questionnaire is received it is marked-off as USI ñ04ò and when this questionnaire is captured and verified then the USI code is changed to ñ07ò. Enterprises are handled according to their respective USI codes.

Table 1.6: Description of each unit status indicator (USI) codes

USI code	**Description**
00	Outstanding questionnaires
01	Undelivered questionnaires
02	Enterprise received as observational units (must have no value)
03	Extension after due date
04	Received completed questionnaires
	05 ï Financial statements
05A	Received financial statement from sampling/observational unit
05B	Completed questionnaire from financial statement
05C	Captured questionnaire done from financial statement
05D	Edited questionnaire done from financial statement
06	Checked questionnaires
07	Captured and verified questionnaires
08	Sampling/observational unit under investigation ï supervisor
09	Sampling/observational unit under investigation ï survey statistician
10	Omit the sampling/observational unit for a short period
11	Legal action to be taken. Refusals.
	12 ï Investigation resulted in demographic changes (with value)
12A	Break up
12B	Merger
12C	Sold
12D	Split off
12E	Take over
12F	Liquidated during the survey period
12G	Closed down during the sample period
12H	Untraceable
12I	Foreign business
12J	Classified out of the sample to an industry not covered by current sample
12K	Imputation from other data
	13 ï Zero contribution (no value)
13A	Merger
13B	Sold
13C	Take over
13D	Closed down
13E	Dormant
13F	Finally liquidated
13G	Secondary duplicate
13H	Included in another enterprise in the current sample
22	Unit to be visited

1.3 Imputation and estimation for the Annual Financial Statistics survey

Estimation can be described as the calculated approximation of a result which is usable even if input data may be incomplete, uncertain, or noisy. Särnada, Swensson & Wreman (1992) defines estimators as statistics to produce values that, for most samples, lie near the unknown population quantity that one wishes to estimate. Estimates in the AFS releases are derived using a combination of received data, weight adjustments for nonresponse and imputation of data.

In the AFS survey, the imputation is only performed on size group one enterprises which are completely enumerated, i.e. each enterprise of size group one is automatically surveyed in the AFS. For size group two, three and four, imputations are not performed for non-responding enterprises. These enterprises have weights attached to them to represent other enterprises in the population that were not sampled. These weights are then further adjusted to compensate for nonresponse of each group size. This method however, ignores the data of the sampled enterprises that do not respond. This is a major contributor to variation in the estimates based on the data as the sample size is reduced. For the AFS survey, imputation of missing results for enterprises of group size one is performed by two general methods, historical information method and ratio imputation method.

1.3.1 Historical information method

The preferred and most common imputation method makes use of historical information of the non-responding units and current trends in the economy. If a size group one enterprise has responded in the previous year but not in the current year, then the previous yearὂs data is used for the current year, but adjusted for growth as indicated by the average inflation rate of the economy for the same reference period. This is also referred to the last value carried forward (LVCF) with adjustment for growth.

1.3.2 Ratio imputation method

When historical information is not available, such as in the case of births or newly sampled enterprises, then the ratio imputation method is used. This involves calculating a ratio, using received enterprises, per industry and per variable. This ratio is then applied to the sampled measure of size for the non-responding enterprise to obtain the estimate for that enterprise.

For example, if information from an enterprise in size group one has not been received in the gold mining industry and there is no previous data available for that enterprise, then a ratio is calculated in this specific industry per variable using the responses from the received enterprises in the gold mining industry. This ratio is then applied to give replacement figures for the non-responding enterprise.

To impute for the missing value, we first compute the ratio of an auxiliary variable (interest received) to the turnover (measure of size provided by SARS) for cases with complete data within a stratum (the gold mining industry) and apply this ratio to the turnover variable. This is then used to impute for the missing interest received. Table 1.7 shows that for enterprise number Enxxxxxxx2, the interest received variable is missing.

Table 1.7: Dataset for interest received for the gold mining industry

Enterprise number	**Turnover (Measure of size)**	**Interest received**	**Interest received**
		Before imputation	**After imputation**
Enxxxxxxx1	70	65	65
Enxxxxxxx2	95	missing	87
Enxxxxxxx3	58	52	52
Enxxxxxxx4	63	60	60
Enxxxxxxx5	100	89	89

Let R denote the ratio of interest received to the turnover, then R is computed as follows:

$$R = \frac{65 + 52 + 60 + 89}{70 + 58 + 63 + 100} = 0.914089\,.$$

Note that when computing the ratio, the cases with missing interest received do not contribute. Applying this ratio to the turnover for Enxxxxxxx2, we can impute for the missing interest received as follows:

Enxxxxxxx2, missing interest received = 0.914089 x 95 = 87.

1.4 Practise in other countries

Statistical agencies in other countries conduct similar surveys for similar purposes as the Annual Financial Statistics survey.

The Australian Bureau of Statistics (ABS) conducts an Economy Wide Survey (EWS) which has a similar purpose as the AFS survey. ABS has developed a strategy to divide the enterprises by a set turnover cut-off, i.e. óabove the lineô units and óbelow the lineô units. All employing enterprises and non-employing enterprises óabove the lineô will be given a chance of selection in the EWS collections and estimation will involve using generalized regression to adjust for non-sampled enterprises. For non-employing enterprises óbelow the lineö, a simple data substitution model is applied. http://www.abs.gov.au/ .

Statistics New Zealand also conducts a similar survey, the Annual Enterprise Survey (AES) which provides financial statistics, financial performance and financial position information about industry groups. Estimates are compiled by aggregating data from different sources, including administrative data, using an estimation technique called rate-up or number-raised estimation. Statistics New Zealand is in the process of evaluating their methodology and will be evaluating the ABS approach of using generalized regression estimation as well as administrative registers. http://www.stats.govt.nz/ .

1.5 Annual Financial Statistics 2008

The AFS 2008 release was published on 22 October 2009. The AFS 2008 sample consisted of 27 633 enterprises of which 11 032 (40%) were size group one enterprises.

Table 1.8: Sample and population size per industry per size group for AFS 2008

Industry	**SIC**	**Size group**	**Population size AFS 2008**	**Sample size AFS 2008**
Forestry and fishing	1	1	193	193
	1	2	123	10
	1	3	503	10
	1	4	--	
Mining and quarrying	2	1	198	198
	2	2	203	101
	2	3	189	101
	2	4	402	119
Manufacturing	3	1	2 539	2 539
	3	2	4 831	4 831
	3	3	5 287	1 509
	3	4	18 829	1 857
Electricity, gas and water supply	4	1	30	30
	4	2	37	20
	4	3	59	20
	4	4	226	20
Construction	5	1	846	846
	5	2	2 818	135
	5	3	2 818	110
	5	4	10 234	170
Trade	6	1	3 151	3 151
	6	2	3 735	222
	6	3	15 326	1 028
	6	4	40 910	1 520
Transport, storage and communication	7	1	831	831
	7	2	525	60
	7	3	2 227	93
	7	4	3 971	112
Real estate, activities auxiliary to financial intermediation and other business services, excluding financial intermediation and insurance	8	1	2 389	2 389
	8	2	2 785	110
	8	3	16 994	465
	8	4	44 818	761
Community, social and personal services, excluding government institutions	9	1	855	855
	9	2	597	353
	9	3	7 288	2 232
	9	4	3 519	632
Overall sample size				**27 633**
Size group 1 sample size				**11 032**

Table 1.8 shows the number of enterprises in the population and the number of enterprises sampled per industry per size group in 2008. Size group one enterprises represented 40% of the total sample.

Figure 1.1: Percentage distribution of enterprises per size group for the AFS 2008 population

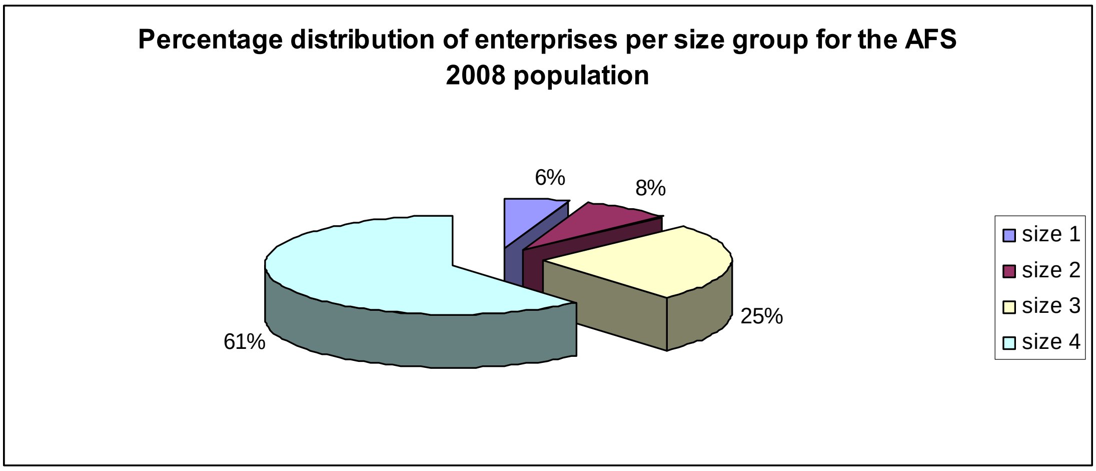
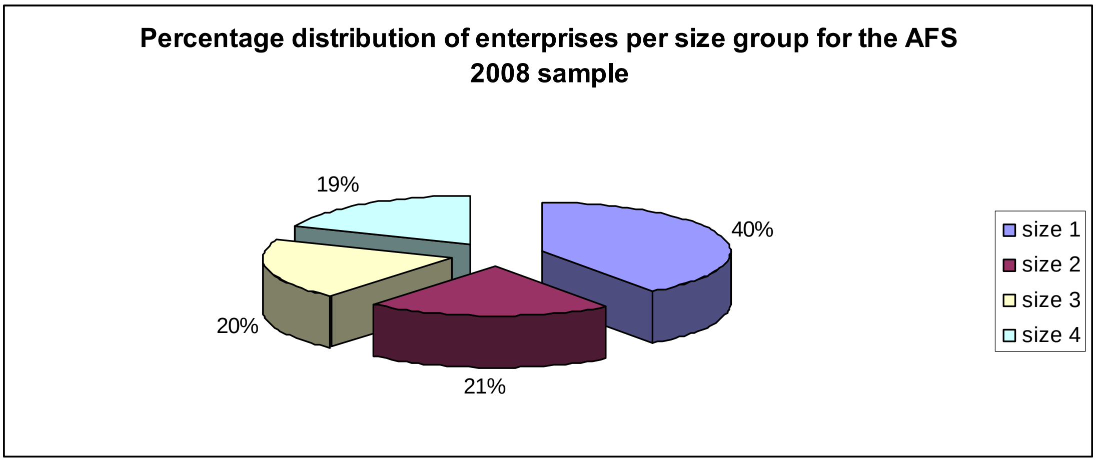


Figure 1.2: Percentage distribution of enterprises per size group for the AFS 2008 sample



Figures 1.1 and 1.2 show the percentage distribution of enterprises per size group for the AFS 2008 population and sample respectively. Although size group one enterprises contribute 8% to the total population for the AFS 2008, it contributes 40% to the AFS 2008 sample.

The USI codes reflect different reaction codes. Table 1.9 shows the grouping of the different USI codes into reaction codes. Group E includes codes that are used during the collection period of the survey. These codes include receiving questionnaires but not capturing or editing them as well cases where enterprises request visits or extensions of the due date. At the end of the collection period no enterprise should have a USI code from group E.

Table 1.9: Grouping of USI codes

Reaction codes	USI codes
A ï Outstanding questionnaires	00
B ï Questionnaires received	05D, 07, 12K
C ï Zero contribution enterprises	12J, 13A, 13B, 13C, 13D, 13E, 13F, 13G, 13H
D ï Other	10, 11, 12A, 12B, 12C, 12D, 12E, 12F, 12G, 12H, 12I
E ï Only used to monitor collection, capturing and editing	01, 02, 03, 04, 05A, 05B, 05C, 06, 08, 09, 22

Table 1.10 shows the number of enterprises that fall into the different reaction groups per size group as well as their respective percentages per reaction code. It can be seen that a total of 14 496 questionnaires were received for AFS 2008.

Zero contribution enterprises include enterprises that where either sold, closed down, liquidated or dormant before the survey period. This group also includes enterprises that are duplicated, included in other enterprises in the survey, foreign businesses and out of scope for the survey (see tables 1.6 and 1.9). These enterprises have zero contribution to the economy and are not imputed for.

ñOtherò includes enterprises that where either sold, closed down or liquidated during the survey period. This group also includes enterprises that were omitted from the sample and those who have refused to complete the questionnaire (see tables 1.6 and 1.9). These enterprises contribute to the economy and are therefore imputed for. This implies that reaction code A - outstanding questionnaires, are not the only group that is imputed for.

Table 1.10: Frequency of different reaction codes per size group for AFS 2008

Reaction codes	**Number of questionnaires**				
	Size group				**Total**
	1	**2**	**3**	**4**	
A - Outstanding questionnaires	772	184	1 124	1 099	3 179
B - Questionnaires received	7 601	2 683	2 550	1 662	14 496
C - Zero contribution Enterprises	1 002	1 649	1 199	1 269	5 119
D - Other	1 657	1 326	695	1 161	4 839
Total	11 032	5 842	5 568	5 191	27 633
Reaction codes	**Percentage per reaction code**				
	Size group				**Total**
	1	**2**	**3**	**4**	
A - Outstanding questionnaires	7%	3%	20%	21%	**12%**
B - Questionnaires received	69%	46%	46%	32%	**52%**
C - Zero contribution Enterprises	9%	28%	22%	24%	**19%**
D - Other	15%	23%	12%	22%	**17%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**

Figure 1.3: Distribution of different reaction codes per size group for AFS 2008

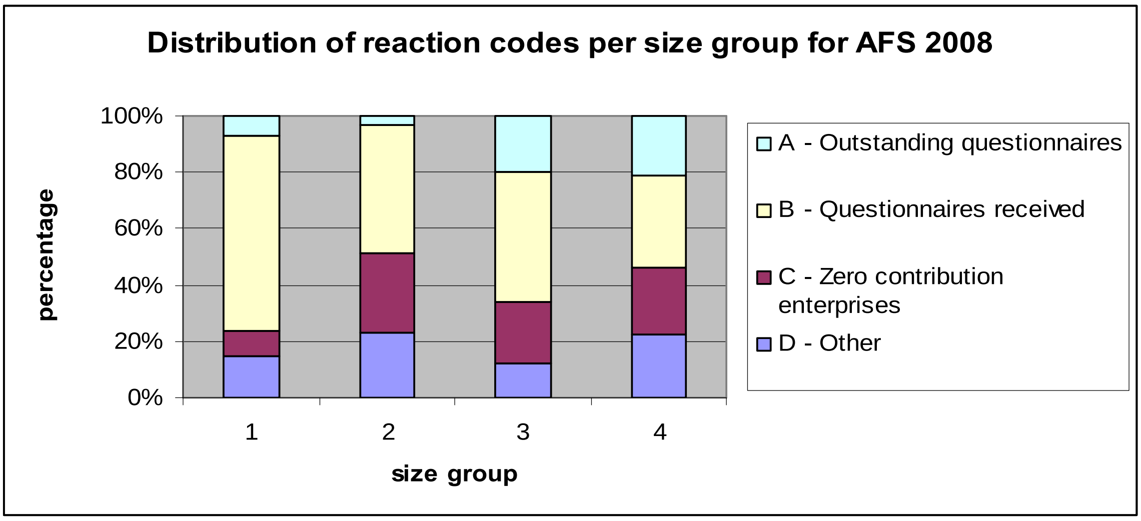
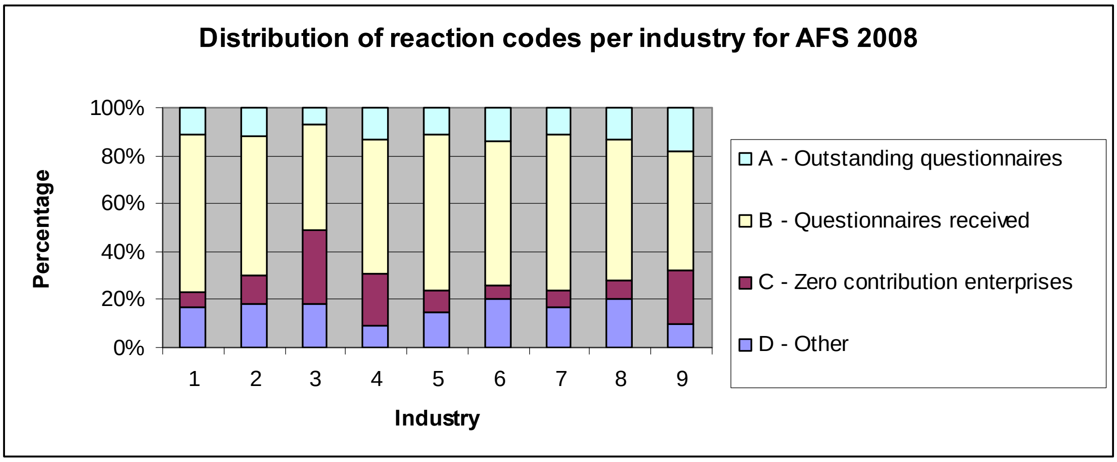


Table 1.11: Frequency of different reaction codes per industry for AFS 2008

Industry	Number of questionnaires				
	Reaction code A	Reaction code B	Reaction code C	Reaction code D	Total
Forestry and fishing	24	140	12	37	213
Mining and quarrying	60	299	60	100	519
Manufacturing	760	4 700	3 280	1 996	10 736
Electricity, gas and water supply	12	50	20	8	90
Construction	140	819	110	192	1 261
Trade	842	3 529	382	1 168	5 921
Transport, storage and communication	122	714	74	186	1 096
Real estate, activities auxiliary to financial intermediation and other business services, excluding financial intermediation and insurance	486	2 208	289	742	3 725
Community, social and personal services, excluding government institutions	733	2 037	892	410	4 072
Total	**3 179**	**14 496**	**5 119**	**4 839**	**27 633**
Industry	**Percentage per reaction code**				
	Reaction code A	**Reaction code B**	**Reaction code C**	**Reaction code D**	**Total**
Forestry and fishing	11%	66%	6%	17%	**100%**
Mining and quarrying	12%	58%	12%	18%	**100%**
Manufacturing	7%	44%	31%	18%	**100%**
Electricity, gas and water supply	13%	56%	22%	9%	**100%**
Construction	11%	65%	9%	15%	**100%**
Trade	14%	60%	6%	20%	**100%**
Transport, storage and communication	11%	65%	7%	17%	**100%**
Real estate, activities auxiliary to financial intermediation and other business services, excluding financial intermediation and insurance	13%	59%	8%	20%	**100%**
Community, social and personal services, excluding government institutions	18%	50%	22%	10%	**100%**
Total	**12%**	**52%**	**19%**	**18%**	**100%**

Figure 1.4: Distribution of different reaction codes per industry for AFS 2008



Weights are used to adjust for size group two, three and four missing enterprises to compensate for outstanding data, i.e. no imputation is attempted for those size groups. Size group one enterprises in groups A and D are imputed. For AFS 2008, a total number of 2 429 enterprises in size group one were imputed for. This contributes 22% of size group one enterprises. Hence, the importance of good imputation methods is crucial in calculating the estimates for the survey.

In this dissertation, an alternative method of imputation will be used (i.e. regression imputation and mean imputation) and the estimates will be recalculated and tested to see how it compares with the method for adjusting for nonresponse that is currently being used by Stats SA.

Chapter 2

Sampling

The need for statistical information is endless in todayôs world as it is required for several decision making and planning criteria for government, businesses etc.. Sometimes it is not feasible to carry out a study to obtain information from the entire population due to the constraints of cost and time. Therefore, in some circumstances it is customary to select a representative subset of the population (sample) and calculate inferences of the population using data obtained from the sample. There are several different sampling selection techniques. This chapter covers the principle steps in a sample survey, and more specifically, concentrates on simple random sampling without replacement and sampling using stratification since the AFS survey uses these sampling methods.

Census refers to a complete enumeration of units, i.e. the entire population whereas sampling refers to partial enumeration, i.e. a subset of the population. The target population for the AFS survey is classed into four group sizes, namely size group i, $i \in \{1, 2, 3, 4\}$; using the cut-off points from the Department of Trade and Industry (see table 1.3). Each enterprise would fall into a group size per industry type according to their measure of size. All size group 1 enterprises are selected for the AFS survey, i.e. a complete enumeration of size group 1 enterprises. For size group 2 ï 4 enterprises, a sample is drawn, i.e. partial enumeration for the size groups. The sampling selection techniques used for the AFS will be discussed later in this chapter.

Cochran (1977) lists the advantages of sampling as compared with complete enumeration:

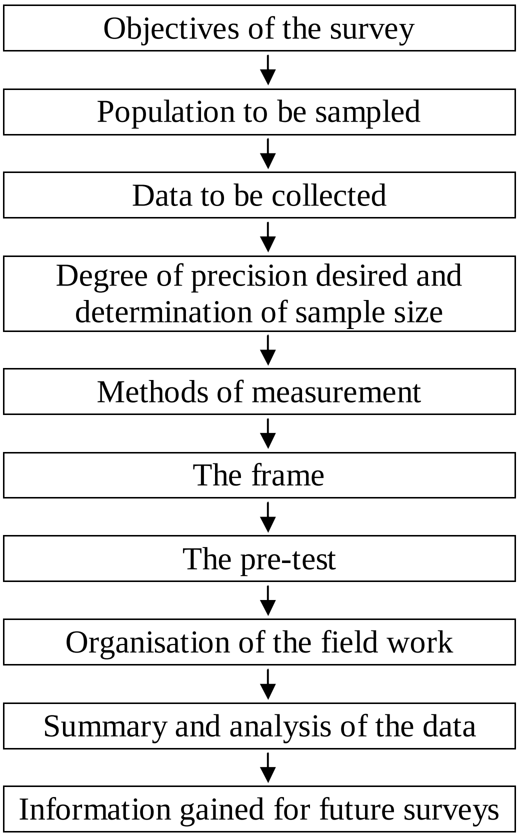
- **Reduced cost**: if data is collected and analysed from a small fraction of the population, the expenditure would be significantly less than if a census is attempted.

- **Greater speed**: data can be collected and analysed more quickly with a sample than a complete count. This is of vital consideration when information is urgently needed.

- **Greater scope**: in certain types of inquiries, highly trained personnel or specialised equipment, limited in availability, must be used to obtain the data. In this case a census would not be practical. Thus surveys have more scope and flexibility regarding the type of information that can be obtained.

- **Greater accuracy**: higher quality personnel can be employed and given intensive training as well as more careful supervision of the field work and processing of the result will produce more accurate results with samples as compared to censuses.

Cochran (1977) explains the principle steps in a sample survey, which is to determine:


Objectives of the survey
Population to be sampled
Data to be collected
Degree of precision desired and determination of sample size
Methods of measurement
The frame
The pre-test
Organisation of the field work
Summary and analysis of the data
Information gained for future surveys

Objective of the survey

It is useful to state the objectives of a survey as it is easy to forget the objectives when one is engrossed in the details of planning.

Population to be sampled

The population to be sampled should coincide with the population regarding which information is required (the target population). For practicability or convenience, the sampled population is sometimes more restricted than the target population. In this case, it should be remembered that conclusions drawn from the sample apply to the sampled population and judgement to the extent to which these conclusions will apply to the target population must depend on other sources of information.

Data to be collected

Questions should be relevant and one should avoid very long questionnaires as this lowers the quality of the answers. Stats SA uses questionnaires as a research instrument for data collection. Questionnaires should be as short as possible without losing any relevant questions. Questions should be concise as well as clear and not ambiguous. Questions can be grouped into two categories, namely open questions and closed questions. Open questions give the respondent a chance to give their own choice of answers while respondents can choose from a set of answers for closed questions. The advantage of open questions is that the researcher is able to get more information while disadvantages include lower response rates due to the time consumption for these questions and it is difficult to put this data into statistical formats. With regard to closed questions, the advantages are it is quick and easy to answer as well as easy to understand. The disadvantages of closed questions include respondents being frustrated if answers liked are not an option and if they feel equally about more than one option.

Degree of precision and determination of sample size

The results of sampled surveys are always subject to some uncertainty as only part of the population is measured. This uncertainty can be reduced by taking larger samples, however, this increases costs and time taken for the survey. Consequently, specifications to the degree of precision in the results, is an important step in the survey. Accuracy refers to the size of deviations from the true mean μ, whereas precision refers to the size of deviations from the mean m obtained by repeated application of the sampling procedure. Cochran (1977) explains that the sample size can be determined by first knowing how accurate the statistician requires the results, e.g. he would be content if the percentage is correct within ± 5%. The sample percentage p is assumed to be normally distributed. In technical terms, p is to lie in the range (P ± 5), except for a 1 in 20 chance. Since p is assumed normally distributed about P, it will lie in the range $(P \pm 2\sigma_p)$, apart from a 1 in 20 chance. Further, $\sigma_p = \sqrt{PQ/n}$. Hence, for our example we will get $2\sqrt{PQ/n} = 5$ or $n = \frac{4PQ}{25}$.

The AFS 2008 sample was sampled at a 3% level of precision. This was found by sampling 14% of enterprises from the formal non-agricultural business sector of the South African economy, excluding financial intermediation, insurance and government institutions. A rough estimate of the size of the sample can be made from the degree of precision required. The method of sampling needs to be planned according to survey requirements.

Methods of measurement

There are many choices of measuring instruments and statistical methods of estimating the population. It is important to have your survey methodology clearly planned for your specific survey. It is good practice to visualise the structure of the final summary tables that will be used for drawing conclusions.

The frame

Before the sample is drawn, the population must be divided into parts called sampling units or units. These units must cover the whole population and must not overlap, that is that every element in the population belongs to only one unit. The collection of all the sampling units is referred to as the sampling frame. Stats SA assigns an enterprise number to all enterprises in the South African economy and uses this as the sampling unit. An enterprise is thus defined as a legal unit or a combination of legal units that includes and directly controls all functions necessary to carry out its production activities.

The pre-test

It is always useful to test out the questionnaire and the field methodology on a small scale before the actual survey. This may reveal troubles which would be easier to fix during the pre-test than when conducting the real survey.

Organisation of the field work

The personnel must receive training on the purpose of the survey and in methods of measurement to be employed in collection of data and there must be adequate supervision of their work. Plans must be made for handling nonresponse.

Summary and analysis of the data

Editing of completed questionnaires assists in amending recording errors and deleting obvious errors. The method of estimation needs to be planned to best suit specific surveys.

Information gained for future surveys

The more information we have initially about the population, the easier it is to draw a sample that will give accurate estimates. All completed samples are potential guides to improved future sampling.

The AFS survey covers the principle steps explained above. International standards and best practices are met in order to obtain a high standard of quality in the statistics produced.

2.1 Sampling designs

Hanurav (1966) explains that a simple finite population ϑ is a population of known number *N* of identifiable units U_1, U_2, é , U_N. A sample *s* from ϑ is an unordered finite sequence of units from ϑ: $s = \{U_{i_1}, U_{i_2}, ..., U_{i_{n(s)}}\}$, $n_{(s)} < \infty$, where $1 \le i_t \le N$ for $1 \le t \le n(s)$. The $i_t$ôs need not necessarily be distinct but the interchange of U_{i_t} and $U_{i_{t'}}$ for $i_t \ne i_{t'}$ results in a new sample. *n*(*s*) is the size of *s*, and *v*(*s*) the number of distinct units of *s*, is the effective size of *s*. While *n*(*s*) can even exceed *N* (because repetitions are allowed), $v(s) \leq N$.

While any specific sample has to be of finite size only, there is no reason, a priori, to restrict one to samples of a fixed size only (i.e. *n*(*s*) = *n*) nor is it obviously justified to restrict to samples of size less than a given number *M* say. Accordingly we define δ, the collection of all possible sample *s* from ϑ as our basic sample space: δ ={*s*}.

Evidently δ contains a countably infinite number of samples and sup *n*(*s*)=∞, $s \in \delta$.

A simple sampling design $D = D(\vartheta, \delta, P)$, briefly called the design *P* is a probability measure *P* defined on δ, $P_s \geq 0$ and $\sum_{s \in \delta} P_i = 1$.

In practice, however, samples are not drawn by listing the *P_s*ôs for all possible samples. Instead, they are drawn by what can be termed as sampling methods. Of particular interest among these sampling methods are the unit drawing mechanisms which are methods of drawing the samples by means of selecting the units from ϑ one by one and with replacement. In its most general form a unit drawing mechanism can be rigorously defined as an algorithm:

$$A = A\{q_1(U_i); q_2(s); q_3(s, U_i)\},$$

where:

1) q_1 is a probability measure on ϑ so that $q_1(U_i) \geq 0$ for $1 \leq i \leq N$ and $\sum_{i=1}^{N} q_1(U_i) = 1$,
2) $q_2(s)$ defined for any sample $s \in \delta$ is a number in (0, 1), $0 \leq q_2(s) \leq 1$ for $s \in \delta$, and
3) $q_3(s, U_i)$ defined only for those *s* for which $q_2(s) \neq 0$ is a probability measure on ϑ: $q_3(s, U_i) \geq 0$ for $1 \leq i \leq N$ if $q_2(s) \neq 0$ and $\sum_{i=1}^{N} q_3(s, U_i) = 1$.

Hanurav (1966) proves that to any given design $D(\vartheta, \delta, P)$ there corresponds a unique unit drawing mechanism $A(q_1, q_2, q_3)$, such that sampling according to *A* results in the design P and conversely.

There are various methods of selecting samples. Mukhhopadhyay (2001), groups these methods into the following categories:

- Simple random sampling with replacement: each unit has an equal chance of being selected. Once a unit has been selected, it is put back into the population thus having an equal chance of being selected once again.

- Simple random sampling without replacement: each unit has an equal chance of being selected. Once a unit has been selected, it is removed from the population thus the unit cannot be selected again.

- Probability proportional to size with replacement sampling: a unit *i* is selected with probability p_i at the *r*th draw and a unit once selected, is returned to the population before the next draw (r = 1, 2, ...). Various items may have different probabilities of being selected for the sample and these probabilities vary in size but are correlated with some measure of size associated with the unit.

- Unequal probability without replacement sampling: a unit *i* is selected at the *r*th draw with probability proportional to some measure of size of that item, denoted by selection probability, $p_i^{(r)}$. Once a unit is selected it is removed from the population before the next selection is made.

- Rejective sampling: draws are made with replacement and with probability { $p_i^{(r)}$ } at the *r*th draw. If all the units turn out distinct, the items selected are taken as the sample; otherwise, the whole sample is rejected and fresh draws are made. In some situations $p_i^{(r)} = p_i \ \forall \ i.$

- Systematic sampling with varying probability: Kish (1965), defines this approach as a selection process that consists of taking every *k*th sampling unit after a random start.

  Rao, Hartley, & Cochran (1962) explain the following sampling procedure. The N units in the population are listed in a random order, their x_t are cumulated and a systematic selection of n units from a random start is then made on the cumulation. A necessary condition is that $np_t \leq 1$ where $p_t = x_t / \sum x_t$. The estimator of the population total Y is $\hat{Y}'' = \sum_{s=1}^{n} \frac{y_s}{P_s}$, where Ps is the probability for the *s*th unit to be in the sample. For this method $P_s = np_s$. Therefore, when the y_t are exactly proportional to the pt, $V(Y\!\!\!\breve{E}'')$ is zero which suggests that considerable reduction in $V(Y\!\!\!\breve{E}'')$ wil result when the y_t are approximately proportional to the x_t.

- Sampling from groups (stratification): the population is divided into L groups (or strata), either at random, or following some suitable procedures and a sample of size n_h is drawn from the hth stratum using any of the above mentioned sampling designs, such that the desired sample size $n = \sum_{h=1}^{L} n_h$ is attained.

In the context of this dissertation, we will focus on simple random sampling without replacement and sampling using stratification as the AFS survey uses these methods of sampling predominantly. The AFS survey is sampled using a stratified random sample design where stratification is based on industry and business turnover. Each enterprise would fall into a group size per industry type according to their measure of size, i.e. their business turnover. All enterprises in the group size representing the large enterprises (size group one) are automatically selected in the sample. This is since size group one enterprises have a large impact on the economy. For size groups two, three and four, enterprises on the business sampling frame are placed into strata, from which a random sample is selected. Each stratum contains two parts of information, namely the industrial sector (primary stratum) to which an enterprise belongs and the group size (secondary stratum) to which an enterprise is assigned. For example, all the size group 2 enterprises in the wholesale trade industry form a stratum. The two parts of information for this stratum would be the industrial sector, i.e. wholesale trade and the group size, i.e. size group 2. Another example of a stratum would be all the size group 4 enterprises in the motor trade industry.

2.2 Simple random sampling without replacement (SRSWOR)

The number of distinct samples of size n that can be drawn from a population of N units using SRSWOR is given by the combinational formula i.e. once a unit is drawn, it is not eligible to be drawn again for that respective sample

$$\binom{N}{n} = {}_N C_n = \frac{N!}{n!(N-n)!} \,.$$

Cochran (1977) defines simple random sampling as a method of selecting n units out of the N population such that every one of the ${}_N C_n$ samples has an equal chance of being chosen. Pathak (1988) views simple random sampling without replacement (SRSWOR) to be commonly perceived as being more practical as well as more efficient than other forms of simple random sampling for sample surveys.

Theorem 2.2.1

Let $\bar{y}_s = \frac{1}{n}\sum_{i \in s} y_i$ be the sample mean based on a sample s of size n by simple random sampling without replacement. Then,

i) $E(\bar{y}_s) = \bar{Y}$

ii) $V(\bar{y}_s) = (1-f)S_y^2 / n$

iii) $Cov(\bar{y}_s, \bar{x}_s) = (1-f)S_{xy} / n$

iv) $\hat{V}(\bar{y}_s) = (1-f)s_y^2 / n$

where $S_y^2 \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{Y})^2$, $S_{xy}^2 \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{X})^2(y_i-\overline{Y})^2$,

$s_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i-\overline{y}_s)^2$ and $f = n/N$.

Proof:

Let ξ be the collection of all possible $\binom{N}{n}$ samples each of which have equal probability of selection$1/\binom{N}{n}$. Let $I_{si} = \begin{cases} 1\text{if } i\in s \\ 1\text{if } i\notin s \end{cases}$

i) $E(\overline{y}_s) = \sum_{s\in\xi}\overline{y}_s / \binom{N}{n}$

$$= \sum_{s\in\xi}\sum_{i=1}^{N} y_i I_{si} / \left[n\binom{N}{n}\right]$$

$$= \sum_{i=1}^{N} y_i\left(\sum_{s\in\xi} I_{si}\right) / \left[n\binom{N}{n}\right]$$

$$= \overline{Y} \text{ Since } \left(\sum_{s\in\xi} I_{si}\right) = \binom{N-1}{n-1}.$$

ii) $E(\overline{y}_s^2) = \sum_{s\in\xi}\overline{y}_s^2 / \binom{N}{n}$

$$= \sum_{s\in\xi}\Bigg\{\sum_{i=1}^{N} y_i I_{si} + \sum_{i\neq}^{N}\sum_{j=1}^{N} y_i y_j I_{si} I_{sj} / \left[n^2\binom{N}{n}\right]\Bigg]\Bigg\}$$

$$= \sum_{i=1}^{N} y_i^2\left(\sum_{s\in\xi} I_{si}\right) + \sum_{i\neq}^{N}\sum_{j=1}^{N} y_i y_j\left(\sum_{s\in\xi} I_{si} I_{sj}\right)\left[n^2\binom{N}{n}\right]$$

$$= \left[\frac{n}{N}\sum_{i=1}^{N} y_i^2 + \frac{n(n-1)}{N(N-1)}\sum_{i\neq}^{N}\sum_{j=1}^{N} y_i y_j\right] / n^2$$

Since $\left(\sum_{s\in\xi} I_{si} I_{sj}\right) = \binom{N-2}{n-2}$ for $i \neq j$.

Thus, $V(\bar{y}_s) = E(\bar{y}_s^2) - \bar{Y}^2 = (1-f)S_y^2 / n$.

iii) Let $d_i = y_i - x_i$ and $\bar{d}_s = \sum_{i\in s} d_i / n$

Then, $V(\bar{d}_s) = V(\bar{y}_s) + V(\bar{x}_s) - 2Cov(\bar{x}_s, \bar{y}_s)$.

Hence,

$$Cov(\bar{x}_s, \bar{y}_s) = \left[V(\bar{d}_s) - V(\bar{y}_s) - V(\bar{x}_s)\right]/2$$

$$= \frac{(1-f)}{2n}\frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i - \bar{Y})^2 + (x_i - \bar{X})^2 - 2\{(y_i - x_i) - (\bar{Y} - \bar{X})\}^2\right]/2$$

$$= \frac{(1-f)}{n} S_{xy}.$$

iv) $E(s_y^2) = \frac{1}{n-1} E\left\{\sum_{i\in s} y_i^2 - n\bar{y}_s^2\right\} = \frac{1}{n-1}\left[E\sum_{i\in s} y_i^2 - n\{V(\bar{y}_s) + \bar{Y}^2\}\right]$

Now noting $E\sum_{i\in s} y_i^2 = \frac{n}{N}\sum_{i=1}^{N} y_i^2$, we can verify the result (iv).

2.3 Stratified sampling

In simple random sampling the variance of $\bar{y}$ depends on the variability of y in the population, i.e. on how much Y_1, Y_2, é ,Y_N vary, as can be seen by $\text{var}(\bar{y}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)S^2$. The variance of $\bar{y}$ can thus be reduced by increasing the sample size n, however, this could be expensive. If the population is very heterogeneous (differs greatly i.e. S^2 is large) and costs limit the sample size n, then it will be difficult to get sufficiently precise estimates using a simple random sample from the population, i.e. difficult to reduce $\text{var}(\bar{y})$. This can be resolved by applying the stratification principle.

For stratified sampling, the population is grouped into non-overlapping sub-populations called strata. A sample of a predetermined size is then selected from each stratum. The selections in the different strata are independent. Any sampling technique can be used to drawn samples from each sample. For example, if a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling. For this dissertation, we will only focus on stratified random sampling, i.e. grouping the population into strata and then selecting a SRSWOR sample from each stratum independently.

Let N_h : total number of units in stratum h;

n_h : number of units sampled from stratum h;

y_{hi} : value obtained from the ith unit sampled from stratum h;

$W_h = N_h / N$: stratum weight;

$f_h = n_h / N_h$: sampling fraction in the stratum.

In order to estimate the population mean per unit, the estimate used in stratified sampling is $\bar{y}_{st}$ (with *st* representing stratification), where

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N}$$

where $N = N_1 + N_2 +$ é $+ N_L$.

Theorem 2.3.1 (Cochran (1977))

If in every stratum the sample estimate $\bar{y}_h$ is an unbiased estimate of $\bar{Y}_h$, the mean of the stratum, then $\bar{y}_{st}$ is an unbiased estimate of the population mean $\bar{Y}$.

Proof:

$$E(\bar{y}_{st}) = E\frac{\sum_{h=1}^{L} N_h \bar{y}_y}{N} = \frac{\sum_{h=1}^{L} N_h \bar{Y}_h}{N}$$

Since the estimates are unbiased in the individual strata. The population mean $\overline{Y}$ may be written as

$$\overline{Y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi}}{N} = \frac{\sum_{h=1}^{L} N_h \overline{Y}_h}{N} = E(\overline{y}_{st})$$

Corollary:

Since $\overline{y}_h$ is an unbiased estimate of $\overline{Y}_h$ for simple random sampling within strata so that, $\overline{y}_{st}$ is an unbiased estimate of $\overline{Y}$ for stratified random sampling.

Theorem 2.3.2 (Adapted from Cochran (1977))

For stratified sampling, the variance of $\overline{y}_{st}$, as an estimate of the population mean $\overline{Y}$, is

$$V(\overline{y}_{st}) = \frac{\sum_{h=1}^{L} N_h^2 V(\overline{y}_h)}{N^2} = \sum_{h=1}^{L} W_h^2 V(\overline{y}_h)$$

where

$$V(\overline{y}_h) = E(\overline{y}_h - \overline{Y}_h)^2 .$$

There are two restrictions on this theorem: (a) $\overline{y}_h$ must be an unbiased estimate of $\overline{Y}_h$ and (b) the samples must be drawn independently in different strata.

Proof:

$$\overline{y}_{st} - \overline{Y} = \frac{\sum_{h=1}^{L} N_h \overline{y}_h}{N} - \frac{\sum_{h=1}^{L} N_h \overline{Y}_h}{N}$$

$$= \frac{\sum_{h=1}^{L} N_h (\overline{y}_h - \overline{Y}_h)}{N}$$

where the sum extends over all strata. The error $(\bar{y}_{st} - \bar{Y})$ in the estimate is now expressed as a weighted mean of the errors of estimation which have been made within the individual strata. Hence

$$(\bar{y}_{st} - \bar{Y})^2 = \frac{\sum_{h=1}^{L} N_h^2 (\bar{y}_h - \bar{Y}_h)^2}{N^2} + \frac{2\sum N_h N_j (\bar{y}_h - \bar{Y}_h)(\bar{y}_j - \bar{Y}_j)}{N^2}$$

where the terms on the right extends over all pairs of strata.

We average over all possible samples. For any cross-product term, we begin by keeping the sample in stratum *h* fixed and average over all samples in stratum *j*. Since sampling is independent in the two strata, the possible samples in stratum *j* will be the same and have the same probabilities, irrespective of which sample has been drawn in stratum *h*. Since $\bar{y}_j$ is assumed unbiased, the average of $(\bar{y}_j - \bar{Y}_j)$ is zero. Thus all cross-product terms will vanish.

The squared terms give

$$V(\bar{y}_{st}) = \frac{\sum_{h=1}^{L} N_h^2 E(\bar{y}_h - \bar{Y}_h)^2}{N^2} = \frac{\sum_{h=1}^{L} N_h^2 V(\bar{y}_h)}{N^2}.$$

Theorem 2.3.3 (Adapted from Cochran (1977))

For stratified random sampling, the variance of the estimate $\bar{y}_{st}$ is

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) \qquad (2.3.1)$$

Proof:

Since $\bar{y}_h$ is an unbiased estimate of $\bar{Y}_h$, theorem 2.3.2 can be applied. Further by theorem 2.2.2, applied to an individual stratum;

$$V(\bar{y}_h) = \frac{S_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

By substitution into the result of theorem 2.3.2, we obtain

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 V(\bar{y}_h) = \frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h)\,.$$

Corollary1:

If the sampling fractions n_h / N_h are negligible in all strata,

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n_h} = \sum_{h=1}^{L}\frac{W_h^2 S_h^2}{n_h}$$

This is the appropriate formula when finite population corrections can be ignored.

Corollary 2:

With proportional allocation, we substitute $n_h = nN_h / N$ in (2.3.1). The variance reduces to

$$V(\bar{y}_{st}) = \sum_{h=1}^{L}\frac{N_h}{N}\frac{S_h^2}{n}\left(\frac{N-n}{n}\right) = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2\,.$$

Corollary 3:

If sampling is proportional and the variances in all strata have the same value, S_w^2, we obtain the result

$$V(\bar{y}_{st}) = \frac{S_w^2}{n}\left(\frac{N-n}{N}\right).$$

2.3.1 Optimum allocation

The sample sizes n_h for stratified samples are chosen by the sampler. It may be selected such that $V(\bar{y}_{st})$ is minimised for a specific cost of taking the sample or to minimise the cost for a specified value of $V(\bar{y}_{st})$.

The cost of the survey can be written as

$$\text{Cost} = C = c_0 + \sum_{h=1}^{L} c_h n_h \quad . \tag{2.3.2}$$

Since within any stratum, the cost is proportional to the size of the sample. The cost per unit c_h may vary for different strata, while the overhead costs are represented by the term c_0.

Theorem 2.3.1.1 (Cochran (1977))

In stratified random sampling with a cost function of $C = c_0 + \sum_{h=1}^{L} c_h n_h$, the variance of the estimated mean $\bar{y}_{st}$ is a minimum when n_h is proportional to $N_h S_h / \sqrt{c_h}$.

Proof:

We need to minimise

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}(1 - f_h) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}$$

subject to

$c_1n_1 + c_2n_2 +$ é $+ c_Ln_L = C$ ï c_0.

Using Lagrange multipliers, we select the n_h and the multiplier λ to minimise

$$V(\bar{y}_{st}) + \lambda\left(\sum_{h=1}^{L} c_h n_h - C + c_0\right) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h} + \lambda\left(c_1 n_1 + c_2 n_2 + \ldots + c_L n_L - C + c_0\right)$$

Differentiating with respect to n_h gives the equation

$$-\frac{W_h^2 S_h^2}{n_h^2} + \lambda c_h = 0\,,\ (h = 1, 2, \text{é}\ \ , L)$$

that is,

$$n_h \sqrt{\lambda} = \frac{W_h S_h}{\sqrt{c_h}} \qquad (2.3.3)$$

Summing over all strata, we obtain

$$n\sqrt{\lambda} = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}} \qquad (2.3.4)$$

The ratio of (2.3.3) to (2.3.4) gives

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L}\left(W_h S_h / \sqrt{c_h}\right)} = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L}\left(N_h S_h / \sqrt{c_h}\right)} \qquad (2.3.5)$$

This theorem leads to the following rules of conduct. In a given stratum, take a larger sample if:

- The stratum is larger,
- The stratum is more variable internally,
- Sampling is cheaper in that stratum,

so that the optimal sample size is selected to get the best estimates at the lowest cost. Equation (2.3.5) gives the n_h in terms of *n*, but we do not yet know what value *n* has. The solution depends on whether the sample is chosen to meet a specific total cost *C* or to give a specified variance *V* for $\bar{y}_{st}$. If the cost is fixed, then substitute the optimum values of n_h in the cost function (2.3.2) and solve for *n*. This gives

$$n = \frac{(C - c_0)\sum_{h=1}^{L}\left(N_h S_h / \sqrt{c_h}\right)}{\sum_{h=1}^{L}\left(N_h S_h \sqrt{c_h}\right)}$$

If *V* is fixed, substitute the optimum n_h in the formula for $V(\bar{y}_{st})$. Then

$$n = \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right) \sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}{V + (1/N) \sum_{h=1}^{L} W_h S_h^2}$$

where $W_h = N_h / N$.

2.3.2 Formation of strata

The main drivers for forming strata are to have units within a stratum to be as similar as possible. Their variances are reduced to the extent that the variation among sampling units within the strata is less than their variation in the entire population. Hence, we strive to increase and maximise the homogeneity of the sampling units within strata. Thus any variable used for stratification requires information to be available on all of the sampling units in the population. If information is only available for a small part of the sampling units then it is generally not useful for stratification.

2.4 Mean square error

Suppose a population parameter (θ) is to be estimated by a sample estimate $(\hat{\theta})$. Mean square error (MSE) is a useful criterion to compare a biased estimate with an unbiased estimate or two estimates that have different amounts of bias. The AFS survey uses this variance estimation as a measure of accuracy between estimates of consecutive surveys.

Theorem 2.4 (Adapted from Cochran (1977))

The mean square error (MSE) is

$$\text{MSE}(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = \text{Var}(\hat{\theta}) + (\text{Bias})^2$$

Proof:

$\mathrm{E}(\hat{\theta}) \neq \theta$ as it has a bias. Let $\mathrm{E}(\hat{\theta}) = m$.

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= E[\hat{\theta} - m + m - \theta]^2 \\
&= E(\hat{\theta} - m)^2 + 2(m-\theta)E(\hat{\theta} - m) + (m-\theta)^2 \\
&= Var(\hat{\theta}) + 2(m-\theta)(m-m) + (m-\theta)^2 \\
&= Var(\hat{\theta}) + (m-\theta)^2 \\
&= Var(\hat{\theta}) + (Bias)^2
\end{aligned}
$$

Since $\mathrm{E}(\hat{\theta}) = m$, let $Bias = m - \theta$ as all unbiased estimators then have $Bias = \theta - \theta = 0$ as expected and $\mathrm{MSE}\left(\hat{\theta}\right)$ = $\mathrm{Var}\left(\hat{\theta}\right)$ in this case.

2.5 Confidence intervals

Särnada, Swensson & Wreman (1992) defines a confidence interval as a random interval $CI(s) = [\theta_L(s), \theta_U(s)]$, where the lower endpoint $\theta_L(s)$and the upper endpoint $\theta_U(s)$ are two given statistics such that $\theta_L(s) \leq \theta_U(s)$ for every *s*. The estimate should lie within the confidence intervals.

There is a need to construct the upper and lower endpoints, $\theta_U(s)$ and$\theta_L(s)$, so that a desired confidence level $1-\alpha$ is attained. For example, a 0,95 confidence interval ($\alpha = 0{,}05$) would mean that the true value of θ would be between $\hat{\theta}_L(s)$ and $\hat{\theta}_U(s)$ 95% of the time, where $\hat{\theta}_L(s)$ and $\hat{\theta}_U(s)$ are the lower and upper limits of the *CI* for θ, calculated based on sample *s*.

Let $\hat{\theta}$be the point estimator for the unknown θ. A confidence interval forθ at the approximate level $1-\alpha$ is computed as

$$\hat{\theta} \pm z_{1-\alpha/2}\left[\hat{V}(\hat{\theta})\right]^{1/2} \tag{2.5.1}$$

where $z_{1-\alpha/2}$ is the constant with probability $\alpha/2$ by the N(0, 1) random variable.

One usually chooses a small value for α. Table 2.1 show the most commonly used values of α.

Table 2.1: Common confidence probability percentages

Confidence probability ($1-\alpha$)	50%	80%	90%	95%	99%
$z_{1-\alpha/2}$	0.67	1.28	1.64	1.96	2.58

The AFS 2008 estimates are calculated within a 95% confidence interval.

If the following two conditions are verified, then the confidence interval calculated from equation (2.5.1) will contain the unknown value θ for an approximate proportion of $1-\alpha$ for repeated samples s drawn with the given designs.

1. The sampling distribution of $\hat{\theta}$ is approximately a normal distribution with mean θ and variance $V(\hat{\theta})$.
2. There exists a consistent variance estimator $\hat{V}(\hat{\theta})$ for $V(\hat{\theta})$.

2.6 Sampling significance on estimation

Although estimates are aggregated at a 1 digit level for the Annual Financial Statistics publication, disaggregated estimates (estimates at a lower digit level) are also made available. Thus estimates are calculated according to the strata sampled. One needs to be familiar with the sampling methodology applied for the AFS to derive the best possible estimates. Imputation for missing data needs to be addressed at the stratum level in order to achieve high quality estimates.

2.6.1 Example of uses of the AFS data

Value-added is the net output per sector. It is obtained by deducing intermediate consumption from the gross output. Value-added is calculated using the AFS data. This is calculated at a disaggregated level for each industry in order to see differences within and across sectors. Disaggregated industry estimates are estimates broken down into more detailed classifications e.g. the mining and quarrying estimates are broken down, showing the estimates for the different types of mining and quarrying such as mining of gold and uranium ore, mining of platinum group metals, mining of copper, quarrying of limestone and limeworks etc. Thus the importance of good imputation of missing data at a stratum level will result in better quality estimates.

Table 2.2 shows the template used to calculate value-added for each classification. National Accounts, a division at Stats SA, provides the sampling specifications to the AFS survey. These specifications are in accordance to the level at which value-added is required as well as GDP levels. Estimates are calculated at the level at which it was sampled.

Table 2.2 shows what data is used from the AFS 2008 survey in the calculation of value-added. The remaining data (shaded in green) is provided from SARS. National Accounts is responsible for calculating value-added as well as GDP at the predetermined levels using both AFS and SARS data.

Table 2.2: Calculation of value-added for platinum group metals for 2008

AFS			
Calc.		**Description**	
no.		**SIC 2424 - Mining of platinum group metals**	R million
	+	Sales of goods	110 968
	+	Income of services rendered	191
	+/-	Changes in inventories: Work in progress **(Calc. no. M)**	
	+/-	Changes in inventories: Finished goods produced **(Calc. no. N)**	
A	=	Output of manufacturing products	
	+	Sales of factored goods	
	-	Purchases of factored goods	
	+/-	Changes in inventories: Factored goods **(Calc. no. O)**	
B	=	Output, trade margins	
	+	Income from mineral rights	0
	+	Income from rental etc (buildings etc)	72
	+	Income from leasing etc (plant etc)	0
	+	Income from leasing etc (vehicles etc)	0
	+	Royalties etc received	0
	+	Other income	594
C	=	Output, miscellaneous services	666
D	=	**OUTPUT at producers prices (Calc. no. A + B + C)**	
	-	Excise and customs duty	0
E	=	**OUTPUT at basic prices**	
	+	Purchases	37 766
	+	Advertising	427
	+	Bank charges	10
	+	Bursaries	0
	+	Containers and packaging materials	1
	+	Insurance premiums	103
	+	Motor vehicles running expenditure	25
	+	Operational leasing and hire of plant, machinery, equipment and vehicles	41
	+	Paper, printing and stationery	13
	+	Postal, courier and telecommunication services	23
	+	Railage and transport-out	268
	+	Rental of land, buildings and other structures and water and electricity services	2 450
	+	Repair and maintenance	475

	+	Royalties, franchise fees, copyright, trade names and patent rights paid	416
	+	Security services	41
	+	Staff training	31
	+	Subcontractors	4 186
	+	Travelling, accommodation and entertainment	74
	+	Other	977
	+/-	Changes in inventories: Raw materials etc. **(Calc. no. L)**	
F	=	**INTERMEDIATE CONSUMPTION**	
G	=	**VALUE ADDED at basic prices (Calc. no. E - F)**	
	+	Salaries and wages	17 082
H	=	**COMPENSATION OF EMPLOYEES**	17 082
	+	Property taxes	2
I	=	**OTHER TAXES ON PRODUCTION**	2
	+	Government subsidies and incentives	0
J	=	**OTHER SUBSIDIES**	0
K	=	**OPERATING SURPLUS (1) (Calc. no. G - H – I + J)**	
	+	Closing values: Raw materials etc.	1 196
	-	Opening values: Raw materials etc.	929
	-	Valuation adjustment **(Calc. no. S)**	
L	=	Changes in inventories: Raw materials etc	
	+	Closing values: Work in progress	5 690
	-	Opening values: Work in progress	4 011
	-	Valuation adjustment **(Calc. no. T)**	
M	=	Changes in inventories: Work in progress	
	+	Closing values: Finished goods	3 184
	-	Opening values: Finished goods	2 724
	-	Valuation adjustment **(Calc. no. U)**	
N	=	Changes in inventories: Finished goods produced	
	+	Closing values: Factored goods	0
	-	Opening values: Factored goods	0
	-	Valuation adjustment **(Calc. no. V)**	
O	=	Changes in inventories: Factored goods	

P	=	**CHANGES IN INVENTORIES (Calc. no. L + M + N + O)**	
	-	Interest received	1 102
	-	Dividends received	552
	-	Profit on foreign exchange as a result of variations in foreign exchange rates and transactions	1 077
	-	Profit on financial and other liabilities: redemption, liquidation and revaluation	16
	-	Profit on financial and other assets: disposal of assets, realisation for cash and revaluation of assets	222
	+	Depreciation provided for during the financial year	4 740
	+	Interest paid	1 157
	+	Losses on foreign exchange as a result of variations in foreign exchange rates or transactions and losses on financial and other liabilities	1 532
	+	Losses on financial and other assets: disposal of assets, realisation for cash and revaluation of assets	49
	+	Mineral rights leased	333
	+	Net profit before providing for company tax and dividends	44 975
	+	Dividends paid or provided for during the financial year	14 925
	+	Valuation adjustments **(Calc. no. W)**	
Q	=	**OPERATING SURPLUS (2)**	
R	=	**Difference between OPERATING SURPLUS (2) and (1) (Calc. no. Q - K)**	
		Production Price Index - Total output	
		Year - Fourth quarter	
		Average	
		Year - Fourth quarter	
		Production Price Index - SA consumption	
		Year - Fourth quarter	
		Average	
		Year - Fourth quarter	
S	=	Valuation adjustment: Raw materials etc **(used: SA consumption)**	
T	=	Valuation adjustment: Work in progress **(used: Total output)**	
U	=	Valuation adjustment: Finished goods **(used: Total output)**	
V	=	Valuation adjustment: Factored goods **(used: Total output)**	
W	=	Total valuation adjustment, inventories **(Calc. no. S + T + U + V)**	

Chapter 3

Missing data

Nonresponse means that *all* the desired data is not obtained from the entire set of elements identified *s*. Usually the set *s* refers to a sample but it could also refer to the entire population in the case of a census. This chapter defines nonresponse and explains procedures to deal with nonresponse including estimation in the presence of nonresponse. The focus will be on historical imputation and ratio imputation as these imputation methods are used for the AFS survey.

Särnada, Swensson, & Wreman (1992) mentions the objective of a survey is to observe the sampled elements with respect to *q* study variables, y_1,é ,$y_{j,}$é y_q. These may typically correspond to *q* items on a questionnaire. Let y_{jk} be the value of the variable y_j for the *k* th element. Let n_s be the size of the sample *s*, i.e. suppose that n_s elements are sampled. Here y_{jk} represents the value of variable $y_{j,}$ for element *k*, $j \in$ (1,2, ..., q), $k \in$ (1, 2, é , n_s).

By full response in a survey it is meant that, after data collection and editing, the available data consists, for every $k \in n_s$, a complete *q*-vector of observed values y_k = (y_{1k},é ,y_{jk},é ,y_{qk}). These values form a data matrix of dimension $n_s \times q$, with no missing values. In all other cases, there is nonresponse present, i.e the data matrix $n_s \times q$ is incomplete after collection and editing.

Nonresponse can be present in one of three forms: non-coverage, unit nonresponse and item nonresponse. Each of these will be discussed briefly.

Non-coverage occurs when the sampling frame omits some of the survey population. This is done either accidentally, for example the omission of certain addresses, or it could be done deliberately, for example the exclusion of remote areas due to cost factors.

Unit nonresponse occurs when no information is collected from a particular sample element. The element k is a unit nonresponse element if the entire vector of y-values, y_k = (y_{1k},é ,y_{jk},é ,y_{qk}), is missing. This could be caused by refusal to respond, a failure to contact the respondent or by misplacing the completed questionnaire.

Lastly, item nonresponse is the failure to collect a particular item of information from a respondent who has supplied other information. The element k is an item nonresponse element if at least one, but not all of q components of the vector y_k = (y_{1k},é ,y_{jk},é ,y_{qk}) are missing. Some reasons for this could be due to the respondent not knowing the answer to the question or for example he/she refusing to answer a sensitive or seemingly irrelevant question.

For the balance of this dissertation, the focus will be on unit nonresponse, y_{k} where $k \in$ (1, 2, é , n_s), thus focusing on the y_{k} type data notation instead of the y_{jk} type data notation as we will be dealing with entire vectors missing and not particular entries in vectors. This is due to the AFS not having the issue of item nonresponse, i.e. enterprises either completely respond to the survey or they do not respond.

Kenward & Carpenter (2007) explains that data are said to be missing completely at random (MCAR) if the probability of being missing is independent of s, $P(k \mid s) = P(k)$. Data are said to be missing at random (MAR) if the conditional distribution of k given the observed data is independent of the unobserved data $P(k \mid s) = P(k \mid s_o)$, where s_o are units with data observed and s_k are units with data missing.

When neither MCAR nor MAR holds the missing data mechanism is said to be missing not at random (MNAR). For this dissertation, focus will be on data missing at random.

For the AFS survey nonresponse and refusals are a bigger concern each year. This results in higher imputation for the survey. Table 3.1 shows that although group A has dropped in percentage from 2006 to 2008, the percentage for group D (which includes refusals) has increased significantly. Both groups A and D (see tables 1.5 and 1.9) need to be accounted for by either imputation or adjustment of weights.

Table 3.1: Frequency of reaction codes for AFS 2006 to 2008

Reaction codes	**Percentage per reaction code - 2006**				
	Size group				**Total**
	1	**2**	**3**	**4**	
A - Outstanding questionnaires	16%	20%	22%	20%	**19%**
B - Questionnaires received	68%	60%	49%	28%	**55%**
C - Zero contribution enterprises	11%	14%	22%	38%	**19%**
D - Other	5%	6%	7%	14%	**7%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**
Reaction codes	**Percentage per reaction code - 2007**				
	Size group				**Total**
	1	**2**	**3**	**4**	
A - Outstanding questionnaires	14%	15%	22%	22%	**17%**
B - Questionnaires received	70%	55%	53%	35%	**57%**
C - Zero contribution enterprises	8%	22%	10%	32%	**17%**
D - Other	8%	8%	15%	11%	**9%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**
Reaction codes	**Percentage per reaction code - 2008**				
	Size group				**Total**
	1	**2**	**3**	**4**	
A - Outstanding questionnaires	7%	3%	20%	21%	**12%**
B - Questionnaires received	69%	46%	46%	32%	**52%**
C - Zero contribution enterprises	9%	28%	22%	25%	**19%**
D - Other	15%	23%	12%	22%	**17%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**

3.1 Dealing with nonresponse

Lynn (1996) explains that there are statistically two problems caused by nonresponse. Firstly, it reduces the sample size which causes an increase in the standard error of the estimates. The second problem caused by nonresponse is introducing biasness to the results. If the nonresponse is not at random, that is if the respondents are often systematically different from the non-respondents then the sample will produce biased estimates.

As an example, consider Tryfos (1966), a cable television company questioning randomly selected subscribers about the quality of the service they receive. It is likely that dissatisfied customers will feel strongly enough about their grievances to take the trouble to respond to the questionnaire with negative comments. Satisfied customers, on the other hand, may not be as motivated to respond to the questionnaire with favourable comments. Consequently, the proportion of unhappy subscribers amongst the respondents may be greater, and the comments on average more negative, leading to distorted or biased results.

Introducing biasness has been extensively discussed by Holt & Elliot (1991). Suppose that the overall population mean $\overline{Y}$ is given by

$$\overline{Y} = W_R \overline{Y}_R + W_M \overline{Y}_M ,$$

i.e. the weighted population mean for the respondents plus the weighted population mean for those whom did not respond.

Table 3.2: Population proportions and means for respondents and non-respondents

	Respondents	Non-respondents
Population proportions	W_R	W_M
Population means	$\overline{Y}_R$	$\overline{Y}_M$

If we assume a simple random sample, only sampled respondents give information and the achieved sample mean is denoted by $\overline{y}_R$ so that

$$E(\overline{y}_R) = \overline{Y}_R = W_R \overline{Y}_R + W_M \overline{Y}_R$$

and

$$\begin{aligned} \text{bias}(\overline{y}_R) &= \overline{Y}_R - \overline{Y} \\ &= (W_R \overline{Y}_R + W_M \overline{Y}_R) - (W_R \overline{Y}_R + W_M \overline{Y}_M) \\ &= W_M \overline{Y}_R - W_M \overline{Y}_M \\ &= W_M (\overline{Y}_R - \overline{Y}_M) . \end{aligned}$$

We note from the bias equation that nonresponse bias will be zero if either of the following conditions are satisfied:

W_M = 0, i.e. no nonresponse.

$\overline{Y}_R$ = $\overline{Y}_M$, i.e. non-respondents on average yield same results as respondents.

Empirical evidence will not support the first condition. Many survey variables are associated with the individual characteristics suggesting unequal response rates for subgroups. Thus, the second condition is unlikely to be met in practice, hence nonresponse bias will result.

To deal with nonresponse certain measures need to be in place. This includes measures taken during the planning stages of the survey, techniques used for data collection, estimation, model assumptions about the response and relations between variables that are used to construct estimators, will all need to be planned taking the reality of nonresponse into account.

Surveys which deal with sensitive issues such as tax evasion and AIDS often have respondents either refusing to participate or giving false responses. In this case, methods that protect identification are a possible solution. This will encourage cooperation as well as truthful responses.

3.1.1 Planning of the survey

Certain precautions can always be taken before the sample is drawn to reduce the nonresponse rate by as much as possible, i.e. as few questions asked as possible to get all the information desired. The questionnaire should be as short and simple to understand as possible. Follow-on questions should be avoided. If the survey requires visiting the respondents, it should be done professionally and scheduled at a time that suits the respondent. The introduction, manner, tone and appearance of the interviewer should be planned as to not cause offence, irritation or dislike to the respondent. In repeated surveys, the frequency in which respondents are asked to participate needs to be considered. Possible rotation of samples or new samples can be used, i.e. for frequent surveys, units being sampled can be replaced with similar units, if they were sampled more than a specified number of times for a specific time period. Response rates are generally negatively affected by a heavy response burden.

3.1.2 Call-backs and follow-ups

It is possible to retrieve responses from non-respondents by contacting them a second time offering incentives, or appealing to their sense of responsibility or duty, could help persuade them to respond. Rewards, appeals and incentives, however, should be used with caution as they may alter what would have been the first-time response, for example, from a desire to please the donor.

If all those whom had not responded, now do so then the problem is solved. The sample averages based on all responses received are unbiased estimators of the population mean. The second time responses have no other effect than delaying the collection of the sample information.

It is, however, not likely to have all the original non-respondents responding in the second round. It may be possible to contact them a third time, perhaps with an even better appeal or incentive. However, again, this must be done with caution to prevent distorted responses. If there are still some respondents that do not respond to the third attempt, a fourth appeal should be considered and even a fifth, and so on.

The additional attempts to receiving responses are time-consuming and add to the cost of the survey. It also increases the likelihood of distorted responses. As a result, the number of re-contacts is usually limited and once all attempts have been exhausted, it is still likely to have some sampled elements who have not responded.

A different approach would be to take a subsample of the non-respondents and make every possible effort to get responses from all elements in this sample. This would allow for unbiased estimation in the sample despite the nonresponse of certain elements in the original sample.

Cochran (1977) shows that by letting $N = N_1 + N_2$, where

N = number of individuals in the population

N_1 = number of individuals who responded

N_2 = number of individuals who do not respond

Let a random sample of *n* elements contain n_1 elements from the response stratum and n_2 elements from the nonresponse stratum. If information can be collected on a sample of $u = \frac{n_2}{k}$ elements from the second stratum, an unbiased estimate of the population average for *y* is given by

$$\breve{MF} = \left[n_1\bar{y}_{n1} + n_2\bar{y}_u\right]/\,n$$

And has variance

$$\mathrm{var}(\breve{MF}) = \frac{k-1}{n}W_2S_{y2}^2 + \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2$$

where $W_2 = \frac{N_2}{N}$ and S_{y2}^2 represent respectively, the weight and variance of the second stratum.

Now $E(\breve{MF}) = E_1E_2(\breve{MF})$ where E_2 denotes the n_1 are fixed,

$$= E_1E_2\left[\frac{n_1\bar{y}_{n1} + n_2\bar{y}_u}{n}\right]$$

$$= E_1\left[\frac{1}{n}(E_2(n_1\bar{y}_{n1}) + E_2(n_2\bar{y}_u))\right]$$

$$= E_1\left[\frac{1}{n}(n_1E_2(\bar{y}_{n1}) + n_2E_2(\bar{y}_u))\right]$$

$$= E_1\left[\frac{1}{n}(n_1\bar{y}_{n1} + n_2\bar{y}_{n2})\right]$$

$= E_1\left[\frac{1}{n}(y_{\bullet n1} + y_{\bullet n2})\right]$, where $y_{\bullet ni}$ denotes the total of the n_i in the 1st sampling

$$= E_1(\bar{y}_n)$$
$$= \bar{Y}$$

Now $\operatorname{var}(\breve{M}) = E_1 V_2(\breve{M}) + V_1 E_2(\breve{M})$ (3.1.1)

$$\operatorname{var}(\breve{M}) = E_1 V_2(\breve{M}) + V_1(\bar{y}_n) \quad \text{since } E_2(\breve{M}) = \bar{y}_n$$

$$= E_1 V_2(\breve{M}) + \frac{1}{n}\left(1 - \frac{n}{N}\right) S_y^2 \quad \text{where } S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$$

and

$$E_1 V_2(\breve{M}) = E_1 V_2\left[\frac{n_1 \bar{y}_{n1} + n_2 \bar{y}_u}{n}\right]$$

$$= E_1\left[\frac{n_2^2}{n^2} V_2(\bar{y}_u)\right] \quad \text{as } \bar{y}_{n1} \text{ is a constant with respect to } V_2$$

$$= E_1\left[\frac{n_2^2}{n^2}\left\{\frac{1}{u}\left(1 - \frac{u}{n_2}\right) S_{n2}^2\right\}\right] \quad \text{where } S_{n2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n2}\left(y_i - \bar{y}_{n2}\right)^2$$

$$= E\left\{\left[\frac{n_2^2}{n^2}\frac{1}{n_2}\left(\frac{n_2}{u} - 1\right)\right] S_{n2}^2\right\}$$

$$= E_1\left\{\frac{n_2}{n^2}(k-1) S_{n2}^2\right\} \quad \text{, since } k = \frac{n_2}{u}$$

$$= \frac{k-1}{n^2} E_1(n_2 S_{n2}^2)$$

Now given that the sample of n gives n_2 non-respondents

$$E_2\left(S_{n2}^2\right) = E_2\left(\frac{1}{n_2 - 1}\sum_{i=1}^{n2}\left(y_i - \bar{y}_{n2}\right)^2\right)$$

$$= \frac{1}{N_2 - 1}\sum_{i=1}^{N2}\left(Y_i - \bar{Y}_{N2}\right)^2 = S_{y2}^2$$

So that

$$E\left(n_2 S_{n2}^2\right) = E_1\left(E_2\left\{n_2 S_{n2}^2 \middle| n_2\right\}\right)$$

$$= E_1\left(n_2 E_2\left\{S_{n2}^2 \middle| n_2\right\}\right)$$

$$= E_1\left(n_2 S_{y2}^2\right)$$

$$= S_{y2}^2 E_1(n_2)$$

$$= S_{y2}^2 (n)\left(\frac{N_2}{N}\right) = S_{y2}^2 n W_2$$

Thus $E_1 V_2(\breve{M}) = \frac{k-1}{n^2} n \left(\frac{N_2}{N}\right) S_{y2}^2$ and finally from (3.1.1)

$$\mathrm{var}\left(\breve{M}\right) = \frac{k-1}{n} W_2 S_{y2}^2 + \left(\frac{1}{n} - \frac{1}{N}\right) S_y^y$$

where $S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2$ for the whole population

and $S_{y2}^2 = \frac{1}{N_2 - 1} \sum_{i=1}^{N2} \left(Y_i - \overline{Y}_{N2}\right)^2$ for the non-respondents.

The first term in $\mathrm{var}(\hat{M})$ is a contribution to the variance (i.e. inadequacy in prediction) due to the fact that not all were responsive. We then chose $u = \frac{n_2}{k}$ elements for personal interviews. The larger the value of *k*, the larger the first term, as one would expect and for *k* = 1, $\mathrm{var}(\hat{M})$ reduces to the result of theorem 2.2.2 for SRSWOR. So the greater the nonresponse, the larger the first term in $\mathrm{var}(\hat{M})$ hence a greater $\mathrm{var}(\hat{M})$, which would thus result in less accurate results.

3.2 Estimation in the presence of nonresponse

Statisticians must often accept that some nonresponse is inevitable and that they need to calculate the best possible estimates from the data that they actually collected. In the planning of a survey, different estimators should be considered and evaluated using alternative, but realistic response models. The aim is to choose an estimator that is bias resistant and has low variance, in order to make the estimator a better predictor of the population statistic.

3.2.1 Weight adjustment for nonresponse

The most common way of dealing with nonresponse is to adjust the design weights based on the assumption that the responding elements represent both responding and non-responding elements. The design weights of the non-respondents are then redistributed amongst the respondents. This is usually done by using a nonresponse adjustment factor that is multiplied by the design weight to produce a nonresponse adjusted weight that automatically corrects for the missing information.

The design weight (W_d) is determined by the sample design,

$$W_d = \frac{N}{n}$$

with N = number of individuals in the population, and

n = number of individuals sampled.

To adjust for nonresponse, Statistics Canada (2003) shows the design weight is multiplied by an adjustment factor to obtain the weight adjusted for nonresponse (W_{nr}).

$$\begin{aligned} W_{nr} &= W_d \times adjustment\,factor \\ &= \frac{N}{n} \times \frac{n}{n_1} \\ &= \frac{N}{n_1} \end{aligned}$$

with the adjustment factor = $\frac{n}{n_1}$, where n_1 is the number of individuals who responded.

Note that data from a census will have a nonresponse weight adjustment, where the original design weight in this case would be equal to one as $N = n$.

3.2.2 Imputation

Imputation is a process used to determine and assign replacement values to resolve problems of missing, invalid or inconsistent data. This allows for a complete data set to be created. Some of the techniques are deductive imputation, mean value imputation, nearest neighbour imputation, multiple imputation, hot-deck and cold-deck imputation and ratio/regression imputation. These methods will be discussed below, focusing on cold-deck imputation as well as ratio imputation as these methods are used for the AFS survey.

Suppose that n_1 of the n values have actually responded and the remaining $n_2 = n - n_1$ values are missing. For simplicity, we assume that $y_{n_1+1},\ldots,y_n$ are missing and need to be imputed.

Deductive imputation

Deductive imputation refers to those instances whereby a missing or inconsistent value can be filled with certainty using a logical conclusion. Usually this method is based upon responses given to other items on the questionnaire. For example, if the values for four items were required as well as the total value for these items and only three values were given together with the total value, then one can deduce the value of the missing item with complete accuracy, eliminating the nonresponse. This imputation method is performed before any other method.

Mean value imputation

Mean value imputation replaces a missing or inconsistent value with the mean value for the imputation class. For example, suppose a questionnaire for a financial survey is missing the value for salaries paid for a company. The missing value can be imputed by the average salaries paid for respondents who correctly reported their salaries paid (the imputation class could consist of respondents in a similar income group). This method imputes the mean of the responding elements to the missing elements. That is for $k \in (n_1+1$, é , $n)$

$$\bar{y}_{n_1} = \frac{1}{n_1}\sum_{k=1}^{n_1} y_k \ ,$$

where $\bar{y}_{n_1}$ denotes the overall respondent mean for the n_1 respondents.

Using the mean value imputation is equivalent to applying the same nonresponse weight adjustment to all respondents in the same imputation class. This uses the assumption that nonresponse is uniform and that non-respondents have similar characteristics to respondents.

Nearest neighbour imputation

This method of imputation selects a donor record based on matching variables. The goal is not necessarily to find a donor that exactly matches that of the recipient but rather to find a donor closest to the recipient in terms of the matching variable, i.e. to find the nearest neighbour.

The missing values $y_{n_1+1},\ldots,y_n$ are replaced by y_{i} where $i \in (1$, é , $n_1)$ from similar responding elements in the sample,

$y_{\mathrm{k}} = y_i$ where $i \in (1$, é , $n_1)$, $k \in (n_1+1$, é , $n)$.

Multiple imputation

Rubin (1987) shows that multiple imputation is a statistical technique which is designed to handle missing data by replacing each missing value with two or more acceptable values representing a distribution of possibilities. It creates a multiple-imputed data set, where each missing datum is replaced by a vector of *m* values. The *m* values are ordered in the sense that the first components of the vectors when substituted for the missing values result in one data set, the second components of the vectors when substituted for the missing values create a second data set and so on.

Hot-deck imputation

Särnadal, Swensson and Wretman (1992) explain that improvement on the mean imputation methods are sought by creating a more authentic variability in the values imputed. In *hot-deck* imputation, missing data is replaced by values selected from respondents in the current survey. The missing values $y_{n_1+1},\ldots,y_n$ are replaced by y_i where $i \in$ (1, é , n_1) from similar responding elements in the sample,

$$y_\mathrm{k} = y_i \text{ where } i \in (1, \text{é} \ , n_1),\ k \in (n_1{+}1, \text{é} \ , n).$$

Cold-deck imputation

Cold-deck imputation uses other sources of data, for example, earlier surveys or historical data. The missing values $y_{n_1+1},\ldots,y_n$ are replaced by y_i where $i \in$ (1, é , n_{s_1}) from a different data source with sample size s_1,

$$y_\mathrm{k} = y_i \text{ where } i \in (1, \text{é} \ , n_{s_1}),\ k \in (n_1{+}1, \text{é} \ , n).$$

For the AFS survey, historical imputation is the first choice of imputation, provided previous data is available. Statistics South Africa (2011), illustrates an example, suppose that turnover for an enterprise Enxxxxxxx2 in stratum 1 is not recorded for year 2008 but historical information on turnover for Enxxxxxxx2 is available for year 2007. Using cold-deck imputation method, we impute the turnover variable for 2008 by using the turnover data for that specific enterprise for 2007. We adjust this figure using a growth factor. The growth factor is calculated by averaging the inflation rate for the economy for the same reference period. The growth rate used for AFS 2008 was 10%.

Table 3.3: Dataset for turnover for stratum 1

Enterprise number	**Turnover**	
	2007	2008
Enxxxxxxx1	9 655	14 500
Enxxxxxxx2	15 500	missing
Enxxxxxxx3	15 000	15 020

Let Y_t^* denote the imputed turnover value for the current period *t*. We compute the imputed value for Enxxxxxxx2 by adding the growth adjustment factor to the turnover value for the previous period, $Y_{t\text{-}12}$ as follows:

$Y_t^* = Y_{t\text{-}12} + (Y_{t\text{-}12} * \textit{growth factor}) = 15\,500 + (15\,500 * 0.10) = 17\,050.$

Ratio/regression imputation

This method uses auxiliary information or responses from other records to create a ratio or regression model that uses the relationship that exists between two or more variables. For example, ratio imputation uses the following model:

$$y_k = \beta x_k + \varepsilon_k$$

where

y_k is the value of the *y* variable for the k^{th} element,

x_k is the value of a related *x* variable for the k^{th} element,

β is the slope of the line (i.e. the change in y_k for one element increase in x_k),

ε_k is assumed to be a random error variable with mean 0 and variance equal to σ^2.

The models assumes that y_k is approximately linearly related to x_k and that the observed values deviate above and below this line by a random amount ε_k.

Values of y_k could be imputed by:

$$\tilde{y}_k = \frac{\bar{y}}{\bar{x}} x_k$$

where

$\tilde{y}_k$ is the imputed value for the variable y for record k,

$\bar{x}$ is the average reported x-value for the imputation class,

$\bar{y}$ is the average reported y-value for the imputation class.

The last value carried forward (LVCF) is a special case of ratio/regression imputation (for the case β = 1) where the value for the current occasion is imputed by using the previous value or adjusting the previous occasionὸs value for growth. It is frequently used for quantitative variables in business survey applications. Little & Rubin (2002) explains that mean imputation can also be regarded as a special case of regression imputation where the predictor variables are dummy indicator variables for the cells within which means are imputed.

Durrant (2005) explains that an advantage of regression imputation is that it can make use of many categorical and numerical variables. The method performs well for numeric data, especially if the variable of interest is strongly related to auxiliary variables. The imputed value, however, is a predicted value either with or without an added on residual and not an actually observed value as in so-called hot deck methods. This could be a problem for imputing certain types of variables such as earning or income variables. Another potential disadvantage of such a parametric approach is that the method may be sensitive to model misspecification of the regression model. If the regression model is not a good fit the predictive power of the model might be poor.

Statistics South Africa (2011), illustrates an example of how ratio imputation is used for the AFS survey. Suppose that the historical data does not exist, then to impute for missing values, we first compute the ratio of an auxiliary variable (interest received) to the turnover (measure of size provided by SARS) for cases with complete data within a stratum and apply this ratio to the turnover variable. This is then used to impute for the missing interest received. Table 3.4 shows that for enterprise numbers Enxxxxxxx2 and Enxxxxxx10 the interest received variables are missing.

Table 3.4: Dataset for interest received for stratum 1

Enterprise number	Turnover (Measure of size)	Interest received	Interest received
		Before imputation	After imputation
Enxxxxxxx1	70	65	65
Enxxxxxxx2	95	missing	88
Enxxxxxxx3	58	52	52
Enxxxxxxx4	63	60	60
Enxxxxxxx5	100	89	89
Enxxxxxxx6	57	53	53
Enxxxxxxx7	61	58	58
Enxxxxxxx8	53	49	49
Enxxxxxxx9	52	49	49
Enxxxxxx10	62	missing	57

Let R denote the ratio of interest received to the turnover, then R is computed as follows:

$$R = \frac{65 + 52 + \ldots\ldots + 49}{70 + 58 + \ldots\ldots + 52} = 0.924125 .$$

Note that when computing the ratio, the cases with missing total interest received do not contribute. Applying this ratio to the turnover for Enxxxxxxx2, we can impute for the missing interest received as follows:

$$Y_{t\,t}^{*} = 0.924125 \times 95 = 88$$

Similarly for Enxxxxxx10,

$$Y_{t}^{*} = 0.924125 \times 62 = 57 .$$

3.3 Annual Financial Statistics 2008

Figure 3.1: Breakdown of enterprises that need to be imputed for AFS 2008

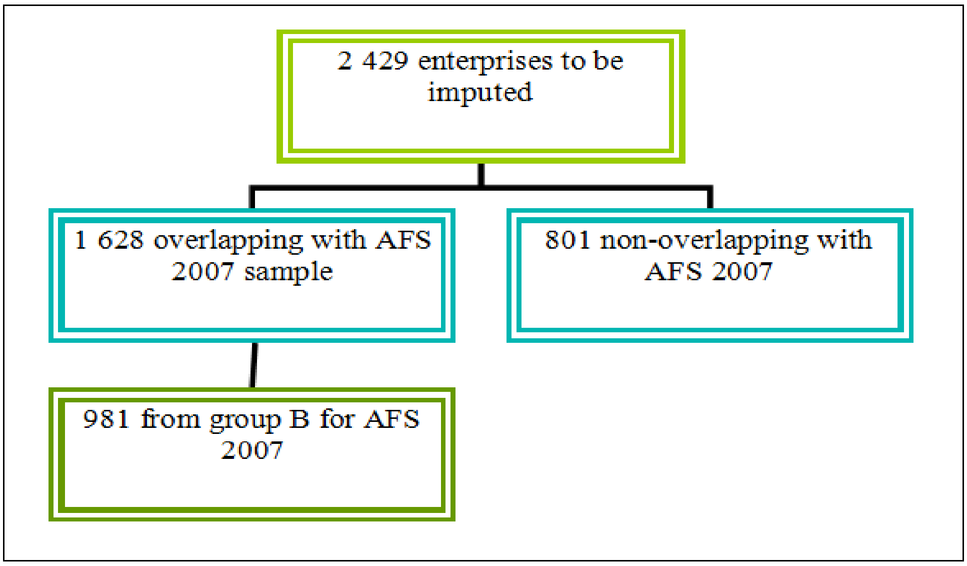

2 429 enterprises to be imputed
1 628 overlapping with AFS 2007 sample
801 non-overlapping with AFS 2007
981 from group B for AFS 2007

Figure 3.1 shows that for AFS 2008 a total number of 2 429 enterprises in group size 1 needed to be imputed. From this total, 1 628 enterprises overlap with the AFS 2007 sample, thus a 67% overlap. This could be due to changes in the company structures resulting in enterprises changing their size group, enterprises closing down as well as new enterprises entering the market. From the overlapping enterprises, 981 enterprises fall in group B for AFS 2007 (group B indicates cases where questionnaires were received, see table 1.9).

Using the current imputation methods used by AFS, historical imputation would be applied to the 981 enterprises that are overlapping with the AFS 2007 sample and have received data for AFS 2007. Hence, 39% of imputation for the group size 1 enterprises are completed using the historical imputation method. The remaining 1 448 enterprises are imputed for using the ratio imputation method.

Chapter 4

Regression imputation

Chapter 3 addresses different types of missing data and methods of handling missing data. Regression imputation is one of the methods of addressing nonresponse. Linear regression analyses the relationship between y on one side and the variables x_1, x_2,é ,x_p on the other side, making it possible to explain y via variables x_1, x_2,é ,x_p. This can be very useful in cases when values of variables x_1, x_2,é ,x_p are easy to obtain whilst this is not the case for the corresponding y values. This chapter investigates the regression imputation method as an imputation technique for the AFS survey.

4.1 Linear regression

Let y be a random variable linearly explained by p variables x_1, x_2, é , x_p. For the i^{th} observed value of y, the linear model becomes

$$\boldsymbol{y_i} = \beta_0 + \beta_1 \boldsymbol{x_{i1}} + \boldsymbol{\ldots} + \beta_{\boldsymbol{p}} \boldsymbol{x_{ip}} + \varepsilon_i \qquad (i = 1, 2, é , n_i), \qquad (4.1.1)$$

where ε_i is the error term, is normally distributed with mean 0 and variance σ^2.

In a matrix form (4.1.1) can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1,p} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or;

$$\boldsymbol{y} = \mathbf{X}\beta + \varepsilon$$

4.1.1 Assumptions

Grob (2003*)* lists the following assumptions that are fundamental for the ordinary linear regression model:

$$\boldsymbol{y} = \mathbf{X}\beta + \varepsilon$$

a. **X** is a non-stochastic $\boldsymbol{n} \times \boldsymbol{p}$ matrix where $p < n$;

b. the matrix **X** has rank *p*, i.e. **X** is of full column rank;

c. the elements of the $\boldsymbol{n} \times 1$ vector $\boldsymbol{y}$ are observable random vectors;

d. the elements of $\boldsymbol{n} \times 1$ vector $\mathcal{E}$ are non-observable random variables such that $\varepsilon \sim (\mathbf{0}, \underline{\sigma^2}\ _n)$, i.e. E(ε) = **0** and *Cov*(ε) = ů$^2\boldsymbol{I}_{\mathrm{n}}$ with ů$^2 > 0$. In other words, on the average, the model is adequate or the ***p*** explanatory variables explain ***y***. The error terms are uncorrelated.

4.1.2 Ordinary least squares estimation

The ordinary least squares principle is presumed to be the best known and most applied method for estimating the regression parameters.

The principal of least squares

If it is assumed that the equation $\boldsymbol{y} = \mathbf{X}\beta^*$ is solvable with respect to β^*, then a solution β^0 satisfies $\left\|\boldsymbol{y} - \mathbf{X}\beta^0\right\|^2 = 0$. On the other hand, if it is assumed that $\boldsymbol{y} = \mathbf{X}\beta^*$ is not solvable then we can determine a vector $\breve{\bar{\beta}}$ such that

$$\left\|\boldsymbol{y} - \mathbf{X}\hat{\beta}\right\|^2 \leq \left\|\boldsymbol{y} - \mathbf{X}\beta^*\right\|^2$$

for every vector β^*. Such a vector is referred to as least squares solution of $\boldsymbol{y} = \mathbf{X}\beta^*$, since the above equation can be re-expressed as

$$\sum_{i=1}^{n}\left(\boldsymbol{y} - \mathbf{X}\hat{\beta}\right)_i^2 \le \sum_{i=1}^{n}\left(\boldsymbol{y} - \mathbf{X}\beta^*\right)_i^2 .$$

If we consider $\varepsilon_i^* = \left(\boldsymbol{y} - \mathbf{X}\beta^*\right)_i$ as the *i*-th residual of the solution β^*, then the sum of squared residuals is minimised for $\beta^* = \hat{\beta}$, so that $\hat{\beta}$ has the smallest sum of squared residuals.

By differentiating the function $\left\|\boldsymbol{y} - \mathbf{X}\beta^*\right\|^2$ with respect to β^* and using

$$\frac{\partial}{\partial \boldsymbol{a}}\boldsymbol{a'Aa} = \left(\boldsymbol{A} + \boldsymbol{A'}\right)\boldsymbol{a} \quad \text{and} \quad \frac{\partial}{\partial \boldsymbol{a}}\boldsymbol{a'b} = \boldsymbol{b}$$

it follows that

$$\frac{\partial}{\partial \beta^*}\left\|\boldsymbol{y} - \mathbf{X}\beta^*\right\|^2 = 2\mathbf{X'X}\beta^* \mathbf{- 2X'}\boldsymbol{y}$$

By equating the right hand side to zero and solving for β^*, we get the solution of $\hat{\beta} = \mathbf{(X'X)}^{-1}\mathbf{X'}\boldsymbol{y}$.

Under the linear regression model with assumptions *a – d in section 3.1.2*, the function

$$\boldsymbol{f}\left(\beta^*\right) = \left\|\boldsymbol{y} - \mathbf{X}\beta^*\right\|^2 = \left(\boldsymbol{y} - \mathbf{X}\beta^*\right)'\left(\boldsymbol{y} - \mathbf{X}\beta^*\right)$$

is minimized for $\beta^* = \hat{\beta}$ where $\hat{\beta} = \mathbf{(X'X)}^{-1}\mathbf{X'}\boldsymbol{y}$.

4.1.3 Analysis of Variance (ANOVA)

The Analysis of Variance (ANOVA) table shows the partitioning of the total variation of a sample into different components, namely, the source, the degree of freedom, the sum of squares, the mean square, the *F* value and the Prob>*F*. Each of these components will be explained below. The source shows the three variations, namely, the model, error and corrected total.

DF is the associated degrees of freedom for each source of variation. JMP, a business unit of SAS (2008), shows that a degree of freedom is subtracted from the total number of non-missing values *N* for each parameter estimate used in a model (corrected total) computation. The total degrees of freedom is partitioned into model and error terms. The model degrees of freedom is the number of parameters *k*, (except for the intercept) used to fit the model. The error degrees of freedom is the difference between the corrected total degrees of freedom and the model degrees of freedom.

The sum of squares shows the amount of variability, with the model sum of squares being the variability explained by the model and the error sum of squares being the variability unexplained by the model. The corrected total, sum of squares shows the total variability in the data. Regression analysis chooses from all possible regression lines by selecting the one for which the sum of squares of the estimated errors is at a minimum. This is referred to as the minimum sum of squared errors (minimum SSE) criterion.

The mean square is the sum of squares divided by its associated degree of freedom. It converts the sum of squares to an average (mean square). The model mean square estimates the variance of the error but only under the hypothesis that the group means are equal. The error mean square estimates the variance of the error term independently of the model mean square and is unconditioned by any model hypothesis.

The *F* value is the model mean square divided by the error mean square. JMP, a business unit of SAS (2008), explains that the *F* value tests the hypothesis that all the regression parameters (except the intercept) are zero. Under this whole-model hypothesis, the two mean squares have the same expectation. If the random errors are normal, then under this hypothesis the values reported in the sum of squares are two independent Chi-squares. The ratio of these two Chi-squares divided by their respective degrees of freedom has a *F*-distribution. If there is a significant effect in the model, the *F* value is higher than expected by chance alone.

Prob>*F* is the probability of obtaining a greater *F* value by chance alone if the specified model fits no better than the overall response mean. Significance probabilities of 0.05 or less are often considered evidence that there is at least one significant regression factor in the model. Large values for the model sum of squares as compared with small values of the error sum of squares will lead to large *F* values and low *p* values, which is what one wants if the goal is to declare that terms in the model are significantly different from zero. It is common to check the *F* test first to make sure that it is significant before delving further into the details of fit.

4.2 Regression diagnostics

Using linear regression models can be very useful if the chosen models are plausible, i.e. there is not substantial indication of inconsistencies or violation of assumptions for the model. Several diagnostic procedures are available to assist with testing the plausibility of the model. It is important to examine the aptness of the model before further analysis based on the model structure can be undertaken.

4.2.1 Selecting independent variables

Suppose we have a dependent variable y which can be explained using linear regression by independent variables from a set { x_1, x_2,é ,x_m}. There are several techniques to select independent variables.

Chatterjee & Hadi (2006), explain the forward selection procedure, the backward elimination procedure and the stepwise method.

Forward selection procedure ï This procedure starts with an equation containing no predictor variables, only a constant term. The first variable included in the equation is the one which has the highest simple correlation with the response variable *Y*. If the regression coefficient of this variable is significantly different from zero it is retained in the equation, and a search for a second variable is made. The variable that enters the equation as the second variable is one which has the highest correlation with *Y*, after *Y* has been adjusted for the effect of the first variable, i.e. the variable significance of the regression coefficient of the second variable is then tested. If the regression coefficient is significant, a search for the third variable is made in the same way. The procedure is terminated when the last variable entering the equation has an insignificant regression coefficient or all the variables are included in the equation. The significance of the regression coefficient of the last variable introduced in the equation is judged by the standard *t*-test computed from the latest equation.

Backward elimination procedure ï This procedure starts with the full equation and successively drops one variable at a time. The variables are dropped on the basis of their contribution to the reduction of error sum of squares. The first variable deleted is the one with the smallest contribution to the reduction of error sum of squares. This is equivalent to deleting the variable which has the smallest *t*-test in the equation. If all *t*-tests are significant, the full set of variables is retained in the equation. Assuming that there are one or more variables that have insignificant *t*-tests, the procedure operates by dropping the variable with the smallest insignificant *t*-test. The procedure is terminated when all the *t*-tests are significant or all variables have been deleted.

Stepwise method ï The stepwise method is essentially a forward selection procedure but with the added proviso that at each stage the possibility of deleting a variable, as in backward elimination, is considered. In this procedure a variable that entered in the earlier stages of selection may be eliminated at later stages. The calculation made for inclusion and deletion of variables are the same as forward selection and backward elimination procedures. Often, different levels of significance are assumed for inclusion and exclusion of variables from the equation.

Mallows' C_m ï This is an alternative measure of total squared error defined as

$$C_m = \left(\frac{SSE_m}{s^2}\right) - (N - 2m)$$

where s^2 is the MSE for the full model and SSE$_m$ is the sum-of-squares error for a model with *m* variables, including the intercept.

AIC criterion ï The Akaikeôs information criterion (AIC) is a relative measure or trade-off between bias and variance in a model. Several competing models can be ranked according to their AIC with the model having the lowest AIC being the best model. AIC is often used for ordinary least square (OLS) method.

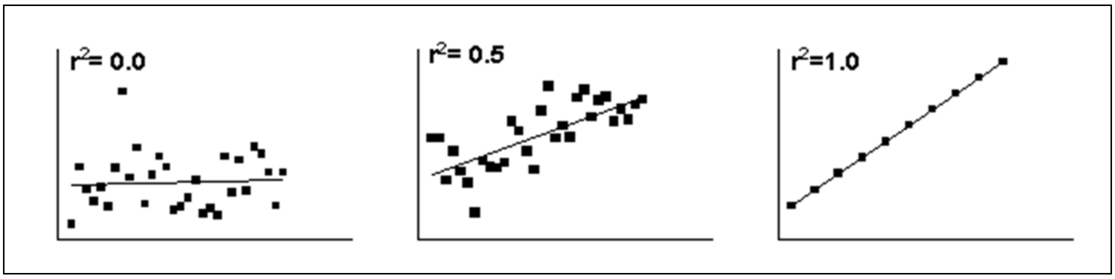
4.2.2 Assessing goodness of fit

The coefficient of determination or R^2 can be seen as a measure on how good the independent variables determine the dependent variables. It can thus be viewed as a measure for the quality of the considered model. It measures the proportion of the variation accounted for by fitting means to each factor level. The remaining variation is attributed to random error. The coefficient of determination, R^2 is calculated as follows:

R^2 = sum of squares (model) / sum of squares (corrected total).

The coefficient of determination, R^2 can give one an idea of the intensity of the fit of the model as R^2 is restricted to the interval [0, 1].

Figure 4.1: Comparison of linear relationship between ***X*** and ***y*** using the coefficient of determination R^2



The value R^2 is a fraction between 0 and 1. Motulsky & Christopoulos (2003) uses figure 4.1 to clearly illustrate that a R^2 value of 0 means, knowing ***X*** does not help you predict *y*. There is no linear relationship between ***X*** and ***y***, and the best-fit line is a horizontal line going through the mean of all *y* values. When R^2 equals 1, all points lie exactly on a straight line with no scatter. Knowing ***X*** lets you predict *y* perfectly.

JMP, a business unit of SAS (2008), explains the adjusted coefficient of determination (adjusted R^2) to be more comparable over models with different numbers of parameters by using the degrees of freedom in its computation. It is a ratio of the mean squares instead of sum of squares and is calculated as follows:

Adjusted R^2 = 1 ï [mean square (error)] / [mean square (corrected total)].

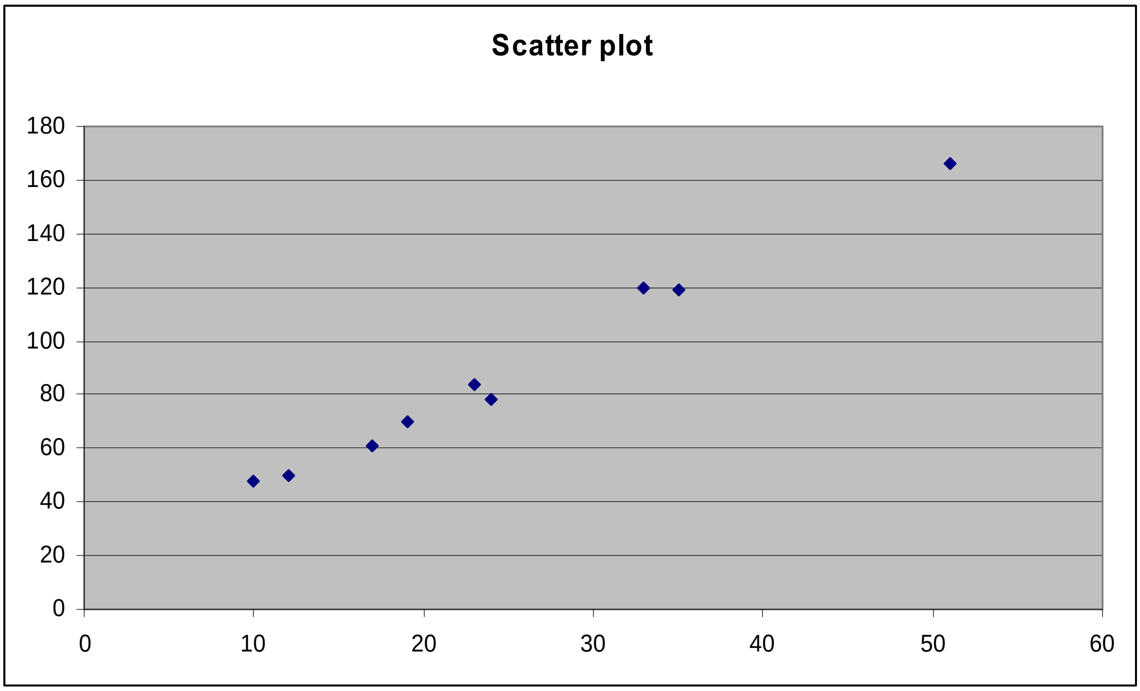
4.3 Prediction in regression

A simple example is provided to explain regression imputation as this application will be used in chapter 5. Table 4.1 shows the dataset used for this example. Ten observations are provided with the *Y* value missing for observation 10. This example shows how a prediction is made for the missing *Y* value using regression.

Table 4.1: Dataset for example on regression

Observation	**X**	**Y**
1	10	48
2	12	50
3	24	78
4	17	61
5	19	70
6	33	120
7	51	166
8	23	84
9	35	119
10	25	é

Figure 4.2: Scatter plot



The data was plotted in a scatter plot. Scatter plots explained by SAS (2007), are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point corresponds to the values of the two variables.

Scatter plots are useful to:

- explore the relationships between two variables;
- locate outlying or unusual values;
- identify possible trends; and
- communicate data analysis results.

A linear regression procedure was run using SAS Enterprise Guide. The following results were achieved.

Table 4.2: Data exploration

Number of Observations Read	10
Number of Observations Used	9
Number of Observations with Missing Values	1

Table 4.2 shows the number of data points used for the exercise. Although 10 observations were read, 9 were used for calculation as the *Y* value for observation 10 is missing.

Table 4.3: Analysis of variance

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	12133	12133	576.13	<.0001
Error	7	147.41378	21.05911		
Corrected Total	8	12280			

The analysis of variance (ANOVA) table provides an analysis of the variability observed in the data and that explained by the regression line. The sum of squares records the sum of squared distances for each source of variation.

The ANOVA table also shows the degrees of freedom (DF) associated with each source of variability. The mean square is the ratio of the sum of squares and the degrees of freedom. For example, the mean square for the error is 147,41378 / 7 = 21,05911. The *F*-value is the ratio of the mean square error for the model and the mean square for the error. This ratio compares the variability explained by the regression line to the variability unexplained by the regression line. SAS (2007) explains that the *F*-value tests whether the slope of the predictor variable is equal to 0. the *p*-value is small (less than 0,05), so you have enough evidence to reject the null hypothesis at the 0.05 significance level. Thus, you can conclude that the simple linear regression model fits the data better than the baseline model. In other words, *X* explains a significant amount of variability of *Y*.

Table 4.4: Measure of fit

Root MSE	4.58902	**R-Square**	0.9880
Dependent Mean	88.44444	**Adj R-Sq**	0.9863
Coeff Var	5.18859		

Table 4.4 provides summary measures of fit for the model. R-square is the coefficient of determination or R^2 value. This value is between 0 and 1 and represents the proportion of variability observed in the data explained by the regression line. In this example, the value is 0,9880, which means that the regression line explains 99% of the total variation in the response values.

The root mean square error (MSE) is the square root of the mean square error. It is an estimate of the standard deviation of the response variable at each value of the predictor variable. The dependent mean is the overall mean of the response variable and the coefficient of variation is the size of the standard deviation relative to the mean. The adjusted R^2 is the R^2 that is adjusted for the number of parameters in the model. This statistic is useful in multiple regression.

Table 4.5: Parameter estimates

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	13.52166	3.47609	3.89	0.0060
X	X	**1**	3.01029	0.12541	24.00	<.0001

Table 4.5 shows the parameter estimates which are the estimated value of the parameters associated with each term in the model as well as the standard errors for each parameter estimate. The *t*-value or *t* statistic is calculated by dividing the parameters by their corresponding standard errors. SAS (2007) explains Pr > $|t|$ is the *p*-value associated with the *t* statistic. It tests whether the parameter associated with each term in the model is different from 0. For this example, the slope for the predictor variable explains a significant portion of variability in the response variable.

Because the estimate of β_0 = 13,52 and β_1 = 3,01, the estimated regression equation is given by Y = 13,52 + 3,01 X.

The missing Y value for observation 10 from table 4.1 can the be calculated using

$$\begin{aligned} Y &= 13{,}52 + 3{,}01\,X \\ &= 13{,}52 + 3{,}01\,(25) \\ &= 89. \end{aligned}$$

This procedure will be applied in the next chapter.

Chapter 5

Application

In this chapter we employ a different imputation method and compare this to the current imputation methods being applied to the Annual Financial Statistics (AFS) survey. The AFS is designed to give information on selected income and expenditure items as well as capital expenditure on new and existing assets, balance sheet items and the carrying value of property, plant and equipment and intangible assets at the end of the financial year for the South African based activities of the enterprise. For this application we will only be looking at 3 variables namely: turnover, purchases and salaries and wages. The same principal can be applied to all the items collected by the survey. We will only look and three industries for this application, namely mining and quarrying, trade and transport, storage and communication. Again, the same principal would apply for all the industries.

5.1 Comparison of imputation methods

Table 5.1 summarises the current method of imputation for the AFS survey and the suggested method of imputation. The data is broken into 3 criterion namely, size group 1 enterprises which did not respond but has received data for the previous period, size group 1 enterprises which did not respond and does not have data for the previous period and size group 2 to 4 enterprises which did not respond. The table shows the method of imputation that would be applied for each criterion for each variable for both the current and suggested methods.

Table 5.1: Comparison of methods to compensate for nonresponse

Criteria	Variable	AFS (current method)	AFS (suggested method)
Size group 1 ï received in previous year	Turnover	LVCF with an inflation factor	Regression imputation
	Purchases	LVCF with an inflation factor	Regression imputation
	Salaries and wages	LVCF with an inflation factor	Regression imputation
Size group 1 ï nonresponse for previous year	Turnover	Use of the BF turnover	Use of the BF turnover
	Purchases	Ratio imputation	Ratio imputation
	Salaries and wages	Ratio imputation	Ratio imputation
Size group 2 - 4	Turnover	Weight adjustment for nonresponse	Mean imputation
	Purchases	Weight adjustment for nonresponse	Mean imputation
	Salaries and wages	Weight adjustment for nonresponse	Mean imputation

For size group 1 enterprises which did not respond but has received data for the previous period, the last value carried forward (LVCF) with an inflation factor (or historical method) is currently used for the AFS survey. The inflation factor is calculating by averaging the inflation indices for the same reference period. For AFS 2008, an inflation factor of 10% was used. Although this is the inflation factor for the current period it does not necessarily mean that it is this uniformed inflation over different industries or different sectors within industries. The new suggested method involves using regression imputation for these cases.

Regression imputation is a technique used to replace nonresponse in data by deriving a linear relationship between a dependent variable AFS 2007 with an independent variable AFS 2008 (see chapter 4 ï Regression imputation). A ñproc regò SAS procedure was used to calculate the estimates using regression imputation. Regression imputation was applied to each stratum of the AFS data (see 4.3 Prediction in regression). Using regression imputation for each variable will allow for each variable trend to be acknowledged. For example, the turnover, purchases and salaries and wages for an enterprise may all not necessarily increase by the same factor. Also, different sector within an industry maybe have different trends thus using regression imputation at a sectorial level will allow for this to be reflected. For example, mining of gold and uranium ore may show a growth in turnover from one reference period to the next, whereas mining of platinum group metals may have a declining turnover trend for the same period. Also, for example, the turnover for the mining of gold and uranium ore sector could show an increase while the interest received for that sector could show a drop for the same reference period. If there was a general drop in interest received for this sector for the enterprises that did respond, the regression imputation technique would apply a similar trend to enterprises in the same sector that did not respond. Each variable trend would be applied to that specific variable per sector for non-responding size group 1 enterprises.

The imputation techniques remain the same for the cases where size group 1 enterprises did not respond and did not have data for the previous period. This group also includes cases where size group 1 enterprises did not respond and were not in the sample for the previous survey. These cases occur when new large enterprises enter the market, smaller enterprises have grown over the period to now become fall in the size group 1 category and mergers and structural changes have resulted in enterprises now falling in the size group 1 category. For this group of enterprises, the BF turnover is used to impute for the turnover variable, this is the measure of size provided by SARS. For the remaining variables, ratio imputation is used (see chapter 3 ï missing data, ratio / regression imputation).

For estimation of size group 2 - 4 enterprises that did not respond, the current method involves adjusting the weights of enterprises in the same sector and size group that have responded to account for those who did not respond. Previous data for these enterprises are not used as there are few overlaps of enterprises for size group 2 - 4. Only size group 1 enterprises are fully enumerated and different samples are chosen for size group 2 - 4 enterprises for each AFS survey. The alternative method uses mean imputation for the size group 2 - 4 enterprises that did not respond. Mean imputation replaces missing data with the arithmetic average of the responded data for the same sector and size group for that specific variable thus maintaining the sample size. Using either the weight adjustment for nonresponse or mean imputation will result in the same final estimation. However, using the weight adjustment for nonresponse approach reduces the sample size hence the precision of the estimates being calculated.

Table 5.2: Summary of statistics for AFS 2008 (current / old) and AFS 2008 (suggested / new)

Variable	Statistics	Mining and quarrying		Trade		Transport, storage and communication	
		AFS 2008 (old)	AFS 2008 (new)	AFS 2008 (old)	AFS 2008 (new)	AFS 2008 (old)	AFS 2008 (new)
Turnover	Estimate (R mill)	330 605	330 492	1 695 210	1 646 116	429 272	422 616
	Standard error	569	251	11 143	5 312	2 915	1 274
	RSE	0,172	0,076	0,657	0,323	0,679	0,302
Purchases	Estimate (R mill)	96 549	96 542	1 328 807	1 355 727	143 168	139 879
	Standard error	301	64	10 193	4 278	1 449	379
	RSE	0,312	0,066	0,767	0,316	1,012	0,271
Salaries and wages	Estimate (R mill)	58 353	58 712	137 547	128 100	59 529	61 163
	Standard error	176	50	1 608	401	727	236
	RSE	0,301	0,085	1,169	0,313	1,222	0,386

Table 5.2 summarises the estimates, standard errors and relative standard errors (RSEs) for the mining and quarrying, trade and transport, storage and communication industries using the old and new imputation methods. The estimates tabulated above are in rand million. The estimates (current / old) for the turnover, purchases and salaries and wages variables for the industries above are the estimates calculated for the South African economy. These estimates (current / old) can be found in the Annual Financial Statistics 2008 publication, Statistics South Africa (2009). The current / old estimates were calculated using the data collected and the current methods to compensate for nonresponse (see table 5.1). The suggested / new

estimates were calculated using the data collected and the suggested methods to compensate for nonresponse (see table 5.1).

The standard errors (SEs) and relative standard errors (RSEs) for each variable per industry are significantly lower when applying the new imputation methods. This implies the range of the confidence interval for the new estimates is narrower hence an increase in the reliability of the estimate. The difference between the current / old estimates and the suggested / new estimates differ per industry and variable. This is due to differences in response rates and trends for each industry and variable. Table 5.1 shows that the suggested / new methods of compensating for nonresponse results in the overall estimates having higher reliability and hence better quality.

5.2 Testing the regression imputation

Regression imputation is only applied to size group 1 enterprises that responded in the previous year (AFS 2007) and did not respond in the current year (AFS 2008). To test the regression imputation method, a population was created using only size group 1 enterprises that had responded for both the current year and the previous year surveys. From this group of enterprises, a randomly selected sample (using the óproc surveyselectô SAS procedure) was removed. Twenty two percent of the data was removed as it was realised that 22% of size group 1 data needed to be imputed, chapter 1 ï table 1.10. Both the old imputation method (LVCF) and the new imputation method (regression) were applied to the dataset with the missing data to create two sets of estimates for this population. This exercise was carried out five times by removing a different sample of enterprises each time and the results were recorded for each scenario. Table 5.3 compares the two estimates with the actual data that was received for the population for each of the five scenarios.

Table 5.3: Comparison of turnover for different scenarios

Scenario	Actual received data	AFS 2008 (old)	AFS 2008 (new)
	R mill		
1	287 599	271 266	296 630
2	333 304	317 274	337 978
3	274 340	256 061	273 949
4	320 939	294 994	326 008
5	321 736	289 895	317 334

In all five scenarios, estimates derived from the new imputation method are closer to the actual data than that derived from the old imputation methods. This again illustrates that trends within each sector per industry are not necessarily uniformed, thus using regression imputation, the individual trends have an influence on the estimates.

From this investigation the evidence suggests that regression imputation will lead to more reliable estimates.

Conclusion

This dissertation shows that although the Annual Financial Statistics (AFS) survey follows international best practice, there are alternative methods to bare in mind when evaluating the methodology used for the survey. The Australian Bureau of Statistics has already implemented the use of regression imputation and Statistics New Zealand is in the process of re-evaluating their methodology and is looking into the implementation of regression imputation for their economic surveys. However, each country has different contributing factors, thus each country has its methodology best suited to work for their needs in accordance to their situation.

In chapter 5 we were able to illustrate the impact of using the different imputation methods for the AFS survey. It was conclusive that a more desired estimate was obtained by using the suggested / new imputation methods. Table 5.3 showed the comparison of turnover for different scenarios. For all scenarios shown, the estimates derived from the new imputation method are closer to the actual data than that derived from the old imputation method. On average there was a 8% improvement in the estimates derived.

Quality is a key area at Statistics South Africa. In order to implement this new methodology, several procedures need to be followed in order to obtain approval. Once the Methodology and Evaluation division approves the suggested changes, the Annual Financial Statistics survey may commence with implementing the proposed changes. Also, other economic surveys within the Economic cluster at Stats SA may benefit from investigating the suggested imputation methods for their surveys and possible implementation thereafter.

One needs to bear in mind that research and investigations on alternative methods of handling missing data should be an on-going exercise. Methods such as multivariate imputation using chained equations (ICE) and inverse probability weighting are examples of areas for future research.

References

Chatterjee, S. & Hadi, A.S. (2006). Regression Analysis by Example, Fourth edition. New Jersey: John Wiley & Sons.

Cochran, W.G. (1977). *Sampling Techniques*, Third edition. New York: John Wiley & Sons.

Durrant, G.B. (2005). Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review. *ESRC National Centre for Research Methods and Southampton Statistical Science Research Institute.* Southampton.

Elliot, D. (1991). *Weighting for non-response – A survey researcher's guide.* Office of Population Census and Surveys (OPCS). Social Survey Division, Statistics Canada.

Grob, J. (2003*). Linear regression.* New York: Springer-Verlag.

Hanurav, T.V. (1966). Some aspects of unified sampling theory, *Sankhya: The Indian Journal of Statistics,* **28**(2), 175-2034

Holt, D. & Elliot, D. (1991). Methods of weighting for unit non-response. *Journal of the Royal Statistical Society. Series D (The Statistician),* **40**(3), 333-342.

JMP, a business unit of SAS (2008). *JMP statistics and graphics guide, Release 8*. North Carolina: SAS Institute Inc.

Kenward, M. and Carpenter, J. (2007) Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16, 199-218.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Little, R.J.A. & Rubin, D.B. (2002). *Statistical analysis with missing data*, Second edition. New Jersey: John Wiley & Sons.

Lynn, P. (1996). Weighting for non-response. *Survey and Statistical Computing*, 205-213.

Motulsky, H.J., & Christopoulos, A. (2003). *Fitting models to biological data using linear and non-linear regression – A practical guide to curve fitting*. San Diego C.A.: GraphPad Software Inc.

Mukhhopadhyay, P. (2001). Topics in Survey Sampling. *Lecture Notes in Statistics*. New York: Springer-Verlag.

Pathak, P.K. (1988). Simple Random Sampling. *Handbook of Statistics.* **6**(1), 97-109. New York: Elsevier Science Publishers B.V.

Rao, J.N.K., Hartley, H.O. & Cochran, W.G. (1962). Procedure of Unequal Probability Sampling without Replacement. *Journal of the Royal Statistical Society. Series B (Methodological),* **24**(2), 482-491.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Särnada, C.E., Swensson, B. & Wreman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

SAS (2007). *Applied analytics using SAS Enterprise Guide Course Notes.* North Carolina: SAS Institute Inc.

Statistics Canada (2003). Survey *methods and practices*, Social survey methods division. Catalogue no. 12-587-XPE, Ottawa.

Statistics South Africa (2009). *Annual Financial Statistics 2008.* Statistical release P0021. South Africa.

Statistics South Africa (2008)^{A}. *Final Sample Findings: Annual Financial Statistics 2008.* Methodology and Standards Division. South Africa.

Statistics South Africa (2011). *Imputation Manual for Economic Statistics Surveys.* Methodology and Standards Division. South Africa.

Statistics South Africa (2006). *Standard Industrial Classification of all Economic Activities*, Sixth edition. Statistics South Africa. South Africa.

Statistics South Africa (2007). *What exactly is the annual financial statistics survey?* Financial Statistics Division. South Africa.

Statistics South Africa (2008)^{B}. *Work programme 2008/09 – 2010/11*. Statistics South Africa. South Africa.

Trade and Industry (2003). *National Small Business Amendment Bill*. Trade and Industry. South Africa.

Tryfos, P. (1966). *Sampling methods for applied research*. New York: John Wiley & Sons.

Websites:

http://www.stats.govt.nz/

http://www.abs.gov.au/

http://www.statssa.gov.za/