The estimation and presentation of standard errors in a survey report

by

René Swanepoel

Submitted in fulfillment of part of the requirements for the degree

Master of Mathematical Statistics

In the Faculty of Science

University of Pretoria

Pretoria

November 2000

Acknowledgements

This study was suggested and initiated by Prof. D J Stoker in his capacity as a consultant to Statistics South Africa. I am also particularly indebted to him for his guidance and support during the study.

I would also like to express my sincere appreciation to Statistics South Africa for making available to me real data sets for the study with the view to develop and test modeling techniques used for the presentation of standard errors in publications. These data sets were: The October Household Surveys (OHS) of 1995, 1996 and 1997, and the Victims of Crime Survey (VOC) of 1998.

Notice

Please note that the three OHS data sets used in the research project differ from the final released OHS data sets in so far the weighting of the data records was based on the adjusted (for growth) 1991 population census data and not on the 1996 population census data. Consequently, the estimates (i.e. estimated values) of population characteristics (such as unemployment rate) appearing in tables in this study, may or will differ from the final released data. For this reason, **all estimates appearing in this study must be considered as privileged and unofficial and may not be quoted in any way whatsoever.**

Note also that the purpose of the study was not to estimate the population characteristics as such, but to model standard errors of the estimated population characteristics with the view to enable readers of survey reports to evaluate the precision of such estimated values.

Contents

Summary

The estimation and presentation of standard errors in a survey report

by

René Swanepoel

Supervisors: Mrs. A Neethling
Prof. D J Stoker

Department of Statistics

Submitted in fulfillment of part of the requirements
for the degree Master of Mathematical Statistics

The vast number of different study variables or population characteristics and the different domains of interest in a survey, make it impractical and almost impossible to calculate and publish standard errors for each estimated value of a population variable or characteristic and each domain individually. Since estimated values are subject to statistical variation (resulting from the probability sampling), standard errors may not be omitted from the survey report. Estimated values can be evaluated only if their precision is known.

The purpose of this research project is to study the feasibility of mathematical modeling to estimate the standard errors of estimated values of population parameters or characteristics in survey data sets and to investigate effective and user-friendly presentation methods of these models in reports. The following data sets were used in the investigation:

- October Household Survey (OHS) 1995 – Workers and Household data set
- OHS 1996 - Workers and Household data set
- OHS 1997 - Workers and Household data set
- Victims of Crime Survey (VOC) 1998

The basic methodology consists of the estimation of standard errors of the statistics considered in the survey for a variety of domains (such as the whole country, provinces, urban/rural areas, population groups, gender and age groups as well as combinations of these). This is done by means of a computer program that takes into consideration the complexity of the different sample designs. The *direct calculated standard errors* were obtained in this way.

Different models are then fitted to the data by means of regression modeling in the search for a suitable standard error model. A function of the direct calculated

standard error value served as the dependent variable and a function of the size of the statistic served as the independent variable. A linear model, equating the natural logarithm of the coefficient of relative variation of a statistic to a linear function of the natural logarithm of the size of the statistic, gave an adequate fit in most of the cases. Well-known tests for the occurrence of outliers were applied in the model fitting procedure. For each observation indicated as an outlier, it was established whether the observation could be deleted legitimately (e.g. when the domain sample size was too small, or the estimate biased). Afterwards the fitting procedure was repeated.

The Australian Bureau of Statistics also uses the above model in similar surveys. They derived this model especially for variables that count people in a specific category. It was found that this model performs equally well when the variable of interest counts households or incidents as in the case of the VOC.

The set of domains considered in the fitting procedure included segregated classes, mixed classes and cross-classes. Thus, the model can be used irrespective of the type of subclass domain. This result makes it possible to approximate standard errors for any type of domain with the same model.

The fitted model, as a mathematical formula, is not a user-friendly presentation method of the precision of estimates. Consequently, user-friendly and effective presentation methods of standard errors are summarized in this report. The suitability of a specific presentation method, however, depends on the extent of the survey and the number of study variables involved.

Opsomming

The estimation and presentation of standard errors in a survey report

deur

René Swanepoel

Studieleiers: Mev. A Neethling
Prof. D J Stoker

Departement Statistiek

Voorgelê ter vervulling van 'n deel van die vereistes vir die graad Magister in Wiskundige Statistiek

Vanweë die groot aantal verskillende populasie parameters en populasie eienskappe in 'n groot steekproefopname, asook die inagneming van al die verskillende subgroepe van belang, is dit bykans 'n onmoontlike taak om vir elke beraamde waarde in die opname 'n standaardfout te bereken. Dit sal ook 'n ontsaglike hoeveelheid ruimte in die publikasie in beslag neem om vir elke beraamde waarde 'n standaardfout in die opname-verslag in te sluit. Die beraamde waardes in die steekproefopname kan egter slegs geëvalueer word indien die presisie van die waardes bekend is. Standaardfoute behoort dus in die opname-verslag ingesluit te word.

Die doel van die navorsingsprojek is om die moontlikheid te ondersoek om deur middel van wiskundige modellering die standaardfoute van die beraamde waardes van die studieveranderlikes van belang in die opname te beraam. Verder moet 'n praktiese, koste effektiewe en gebruikersvriendelike voorstellingsmetode gevind word om die standaardfoute sinvol in die opname-verslag in te sluit. Die data wat in die ondersoek gebruik is, sluit in:

- Oktober Huishoudings Opname (OHO) 1995 – Werkers en Huishoudings datastel
- OHO 1996 – Werkers en Huishoudings datastel
- OHO 1997 – Werkers en Huishoudings datastel
- Die Slagoffers van Kriminele Oortredinge Opname 1998

Die basiese metodologie behels die berekening van standaardfoute van die beraamde populasie parameters of eienskappe vir verskeie subgroepe (soos die RSA, provinsies, landelike / stedelike gebiede, rasgroepe, geslag en ouderdomsgroepe asook kombinasies hiervan) deur middel van 'n rekenaarprogram wat die

kompleksiteit van variansieberaming in die geval van 'n komplekse steekproef in ag neem. Daar word na die standaardfoute wat sodoende verkry word, verwys as *direk berekende standaardfoute.*

Regressie modellering word toegepas op die beraamde waardes en hul direk berekende standaardfoute om sodoende 'n geskikte standaardfout-model te vind. 'n Lineêre funksie met die natuurlike logaritme van die koeffisiënt van relatiewe variasie as afhanklike veranderlike en die natuurlike logaritme van die beraamde waarde as onafhanlike veranderlike, het in die meeste gevalle die beste passings resultate gelewer. Bekende toetse is toegepas om uitskieters te identifiseer waarvoor daar vasgestel moet word of dit geregverdig is (bv. wanneer die subgroep steekproefgrootte te klein is of die beraming sydig is) om die waarnemings uit die datastel te verwyder. Daarna word modelpassing herhaal.

Bogenoemde model word ook deur die Australiese Statistiese Buro (ABS) gebruik in soortgelyke opnames. Hulle het die model afgelei vir 'n veranderlike wat persone in 'n sekere kategorie tel, maar uit die resultate blyk dit dat die model ewe geskik is vir veranderlikes wat huishoudings tel, soos in die geval van die OHO, of gebeurtenisse tel, soos in die geval van die Slagoffers van Kriminele Oortredinge Opname. Die versameling van subgroepe beskou in die ondersoek sluit in kruis-klasse, gesegregeerde klasse en gemengde klasse. Die model kan dus gebruik word ongeag tot watter subklas-tipe die subgroep behoort. Die resultaat maak dit moontlik om standaardfoute vir enige subgroep te modelleer m.b.v. 'n enkele model.

Die model wat op die data gepas is, as 'n wiskundige formule, is nie 'n gebruikersvriendelike voorstellingsmetode van presisie van beramings in die opname nie. Gevolglik vorm die ondersoek en opsomming van effektiewe voorstellings-metodes deel van die navorsingsprojek. Die geskiktheid van 'n voorstellingsmetode hang egter grootliks af van die omvang van die opname en die aantal studieveranderlikes betrokke.

1. Introduction

Addressing the presentation of standard errors is a common problem every survey statistician has to deal with during the compilation of a survey report. The problem is two-fold. Firstly, standard errors of the published survey statistics need to be calculated, and then presented in a simple, comprehensive and cost effective way in the publication.

All estimates of population parameters or characteristics derived from sample survey data are subject to errors. These errors are divided into two categories, viz. sampling- and non-sampling errors. Sampling errors refer to the probabilistic nature of a sample and can be explained as the error made when the sample used for the specific survey is only one of a large number of possible samples of the same size and sample design that could have been selected. Non-sampling errors refer to response differences, definitional difficulties, respondent inability to recall information, etc.

It is impractical to include in a survey report standard errors for each and every statistic, for each and every domain of interest and, taking into account the time absorbency of these complex calculation procedures, it would be an impossible task.

The easiest approach would be to omit standard errors totally from the publication, but there are certain criteria to which published results, subject to the above mentioned errors, have to conform (Gonzalez, Ogus, Shapiro and Tepping; 1975):

a) The user must be informed of the different errors that play a role and the limiting effects of these errors on the results. An explanation of how to interpret standard errors and confidence intervals should be included.

b) The implications of the sample design on the various sources of error must be clearly indicated, e.g. what the effect could be of an old or incomplete sampling frame on the data.

c) If missing data was imputed, it should be mentioned as well as the imputation method that was used and the implications this could have on the results.

d) Standard errors should be displayed in an organized manner and be thoroughly explained.

e) If the results in a survey report are subject to large survey errors, users should be adequately warned against lack of reliability of such data.

Alternatively indirect methods can be used, i.e. modeling standard errors of the survey estimates instead of calculating standard errors for each statistic individually.

The purpose of this research project is to investigate and introduce alternative methods to generate and present standard errors in an efficient way in a survey report. Different aspects that play a role in choosing an acceptable model to approximate standard errors are investigated. Also included, among other factors, are the influence of the size of the subclass to which the estimate belongs, in the model, the effect of the population parameter being estimated in the model and the possible influence that cross-class, segregated class or mixed class domains could have on the model.

1.1 Background: Complex Sampling and Variance Estimation

Before any attention can be given to the finding of a suitable model as the solution for the above-mentioned problem, it is necessary to understand what sampling method was used in the survey. It is also necessary to understand the format of the gathered data and to be informed about the possible consequences that the chosen sampling method could have on the estimation of variances in the survey. The next paragraphs explain the background and the actions taken to arrive at Appendix B, an example of the data used in the modeling procedure to find a suitable model for the estimation of standard errors.

Annually in October, a comprehensive survey called the October Household Survey (OHS), is conducted by Statistics South Africa. This survey gathers information on the employment and unemployment in South Africa and on the general living standards of the people in South Africa. It includes aspects like "the main lighting source" used in a household, whether the household has "running water from a tap " or not and if the household has a " toilet on site" or " toilet off-site", etc.

Although the exact sampling scheme of each OHS differs year by year, the general sampling approach for these surveys can be summarized as follows. Complex sampling is used. Firstly, the whole country is stratified according to a number of stratification variables. From each stratum, Enumerator Areas (EA's)* are drawn as the primary sampling units (PSU's)* and from each drawn EA, a number of households are drawn as the ultimate sampling units (USU's)* (Neethling, Stoker and Eiselen; November 1997).

Using a complex sampling scheme largely complicates variance estimation in a survey, but the advantages of Complex Sampling (CS) make it much more desirable than Simple Random Sampling (SRS), if not the only feasible approach (Neethling, Stoker and Eiselen; November 1997):

- CS makes a step-by-step design of the sample possible
- CS is more economical and practical
- CS guarantees a sample more representative of the population
- CS does not require a complete sampling frame of the population elements.

Another factor that plays an important role in the sampling procedure is whether sampling is done *with* or *without replacement**. In order to avoid multiple drawing of the same sampling unit at any stage, sampling for the OHS is done without replacement (WOR). This adds further difficulty towards the estimation of variances, but Kish's approach (Kish; 1965) provides a simplifying solution in practice. Kish advises the use of WOR sampling in all sampling stages, but to use the formula of with replacement (WR) sampling for variance estimation. Although this approach tends to overestimate the true variance of an estimator, the extent of overestimation is small in general (Neethling, Stoker and Eiselen; November 1997).

After sampling and data gathering have been completed, the data must be processed before starting with the estimation of population parameters and their related variances. Data processing is done by running consecutive SAS programs. The programs were developed by D. J. Stoker, a consultant to Statistics South Africa and have the main purpose of estimating the standard errors and the coefficients of relative variation in a complex sample.

---

* See Glossary

The data is usually categorized according to the following categories. The numbers given below, indicate the categories in the programs.

Table 1: Data categories considered

Provinces:	1 – Western Cape 2 – Eastern Cape 3 – Northern Cape 4 – Free State 5 – Kwazulu / Natal 6 – North West 7 – Gauteng 8 – Mpumalanga 9 – Northern Province
Urban / Rural :	1 – Urban 2 – Rural
EA Type:	1 – Urban formal 2 – Urban informal 3 – Tribal 4 – Commercial farms 5 – Other non-urban
Race:	1 – African / Black 2 – Coloured 3 – Indian 4 – White
Gender:	1 – Male 2 – Female
Age group:	1 – Age 15 to 30 years 2 – Age 31 to 45 years 3 – Age 46 to 65 years
Education group:	1 – None 2 – Some primary 3 – Primary completed 4 – Some secondary 5 – Std 10 completed 6 – Tertiary

All the above categories give rise to a data set consisting of many different subgroups, referred to as domain subclasses (see page 15). Each domain subclass can be classified further as a *cross-class, segregated class* or *mixed class**.

Apart from the above categories, the data is also categorized according to the different subclasses of the study variable of interest. The study variables considered for the research project are:

* See Glossary

- the total number of unemployed people and of employed people according to the *official** or *strict** definition and the *expanded** definition of unemployment;
- the number of *economically active** people in South Africa;
- the number of households with different dwelling-types, e.g. ranging from formal dwellings like a house on a separate stand or a flat in a block of flats to informal housing or shacks and traditional dwellings;
- the number of households with different main water sources, e.g. ranging from running tap water in the dwelling to a borehole either on site or communal, to a stream or spring;
- the number of households with different main lighting sources, e.g. ranging from electricity to wood or candles
- and the number of households with different sanitation facilities, e.g. ranging from a flush toilet in the dwelling to a chemical toilet or to a pit latrine.

The final program does the actual computation to calculate the estimated values of the following population parameters: totals, ratios, standard errors, coefficients of relative variation. Other values of interest like the *design effect** and the upper and lower boundaries of the 95% confidence intervals of the estimated values are also calculated.

This program operates on the principle of repeating the same set of instructions by means of a "do-loop" macro for all the observations that satisfy the specified categorical criteria of the study variable of interest. If the specified condition is true, the study variable y is assigned the value 1 and if not, it is assigned 0. The auxiliary study variable x is used to specify the subgroup under consideration. Note that $y = 0$ when $x = 0$ and if $y = 1$, then $x = 1$ as well. E.g. if the study variable of interest is the number of households with a toilet off-site, and this value is being estimated for each province separately (i.e. province is considered as a domain), the categorical criteria can be expressed in the program as follows:

```
%Do i = 1 %to 9;                */ Number of provinces */
   If prov = &i. then x = 1; else x = 0;
   If x = 1 and 1<=toiloff<=5 then y = 1; else y = 0;
```

where "toiloff" indicates the type of offsite toilet facility used.

The categorical criteria change each time after the program has completed the calculations for a specified range of categories. This process is repeated until estimates are obtained for all domain subclasses of the study variable of interest for which estimates are published.

After each completion of the set of instructions in the "do-loop" macro, the output is written in text format to the file allocated for the specified study variable of interest. Each record is stored on a separate line by making use of the statement "*file out ls=500 mod;*" where *ls = 500* is used when the line size becomes too big. This text file can then be imported and edited in other available software packages, e.g. in MICROSOFT EXCEL.

---

* See Glossary

The following formulas and notation are used in the program for the calculation of population parameters and characteristics. These values are estimated based on the assumption: Let the weight attached to a record (h,i,j) be denoted by w_{hij}, where h is the stratum index, i the PSU index, i.e. the i-th drawn EA, j the USU index, i.e. the j-th drawn household and N the number of population elements, then the weights are such that $\sum_h \sum_i \sum_j w_{hij} = N$.

The estimators that are used in the program for the estimation of population parameters or characteristics are (Neethling, Stoker and Eiselen; November 1997):

Estimator of a total: $$\hat{Y} = \sum_h \hat{Y}_h = \sum_h \sum_i \hat{Y}_{hi} = \sum_h \sum_i \sum_j w_{hij} \hat{Y}_{hij} \quad (1)$$

and $$\hat{X} = \sum_h \hat{X}_h = \sum_h \sum_i \hat{X}_{hi} = \sum_h \sum_i \sum_j w_{hij} \hat{X}_{hij} \quad (2)$$

The ratio estimator:

$$\hat{R} = \frac{\sum_h \sum_i \sum_j w_{hij} \hat{Y}_{hij}}{\sum_h \sum_i \sum_j w_{hij} \hat{X}_{hij}} = \frac{\sum_h \sum_i \hat{Y}_{hi}}{\sum_h \sum_i \hat{X}_{hi}} = \frac{\sum_h \hat{Y}_h}{\sum_h \hat{X}_h} = \frac{\hat{Y}}{\hat{X}} \quad (3)$$

The variance estimator for the estimation of the variance of a population total used in the program is given by (adopting the notation of Verma (Verma; 1982)):

$$\text{var}(\hat{Y}) = \sum_h \frac{a_h}{a_h - 1} \left[\sum_i \hat{Y}_{hi}^{\,2} - \frac{\hat{Y}_h^{\,2}}{a_h} \right] \quad (4)$$

where a_h is the number of sampling units in the h-th stratum.
The alternative formula of Verma for estimating the variance of a ratio estimate:

$$\text{var}(\hat{R}) \approx \frac{1}{\hat{X}^2} \sum_h \frac{a_h}{a_h - 1} \left[\sum_i z_{hi}^{\,2} - \frac{z_h^{\,2}}{a_h} \right] \quad (5)$$

where $z_{hi} = \hat{Y}_{hi} - \hat{R}\hat{X}_{hi}$; $z_h = \sum_i z_{hi} = \hat{Y}_h - \hat{R}\hat{X}_h$ and $\sum_h z_h = \hat{Y} - \hat{R}\hat{X} = 0$ for $\hat{R} = \frac{\hat{Y}}{\hat{X}}$.
The latter serves as a test for correctness of calculations (Stoker; January 1999).
Formulas (4) and (5) are used in practice for complex sampling (Cochran; 177: p. 307).

The estimated standard error:

$$se(\hat{Y}) = \sqrt{\text{var}(\hat{Y})} \quad (6)$$

The estimated coefficient of relative variation:

$$cv(\hat{Y}) = \frac{se(\hat{Y})}{\hat{Y}} \qquad (7)$$

Ratio estimates are biased. In order to limit the bias in the ratio estimate, it is necessary to keep the estimated coefficient of relative variation of $\hat{X}$: $cv(\hat{X}) \leq 0.1$ (or ≤ 0.15), where $\hat{X}$ is the denominator in the ratio estimate (equation (3)) (Stoker; January 1999).

The design factor $(deft)$ is defined as the square root of the design effect $(deff)$* ($deft = \sqrt{deff}$) and it indicates whether the sample size needs to be larger than the sample under SRS in order to obtain the same precision. Although a larger sample size is required when $deft > 1$, the cost per unit is lower under CS and it is more convenient than SRS. The design effect is not explicitly used in the regression modeling, but it is implicitly included in the obtained standard error estimates as can be seen on page 20.

Returning to the example, Appendix - B shows the 1997 OHS Workers data set after the above SAS programs were run and the obtained output was edited in EXCEL in order to exclude the columns in the EXCEL sheet that would not be used in the modeling procedure. This data set (Appendix - B) will now be used in SAS INSIGHT for regression modeling in order to find a suitable standard error model.

To illustrate the output of the SAS programs (as given in Appendix - B), the data record for the number of unemployed men in the Western Cape (from the 1997 OHS Workers data set) is explained. Take note that the estimated values differ from the final released OHS data sets and may not be quoted or used. Thus, the only purpose of this example is to illustrate the output of the programs and the estimated values may not be regarded as actual information.

At the top of page B–3 of Appendix - B, the data record for unemployed men (GENDER = 1) in the Western Cape (PROV = 1) can be found. The sample size of economic active men in the Western Cape is (N) 2977 and the number of unemployed men in the sample for this province (n) is 246. This gives an estimated number of unemployed men in the Western Cape for 1997 (MSWY) of 81091, the estimated number of economic active men in this province as (MSWX) 910511 and the estimated unemployment rate for men in the Western Cape (R) as $\frac{81091}{910511} = 0.0891$.

The next three columns contain the estimated standard errors directly calculated by the above SAS programs: $se(\hat{R}) = 0.0079$ (SE-R), $se(\hat{Y}) = 7394$ (SE-WY) and $se(\hat{X}) = 19998$ (SE-WX). The last three columns contain the estimated coefficients of relative variation: $cv(\hat{R}) = 0.089$ (CV-R), $cv(\hat{Y}) = 0.0912$ (CV-WY) and $cv(\hat{X}) = 0.022$ (CV-WX).

Once the data set is edited and imported into SAS INSIGHT, new variables can be created to use in the modeling procedure. Examples are the natural logarithm of the estimated total $\hat{Y}$ $(\ln(\hat{Y}))$, to be used as the independent variable in the model, and the natural logarithm of the coefficient of relative variation of $\hat{Y}$ $\left(\ln(cv(\hat{Y}))\right)$, to be used as the dependent variable in the model, or any other variable needed in the search for the best suitable standard error model for the data set considered.

* See Glossary

2. Different Proposed Models

In the view of having no better alternative than to make use of indirect methods like mathematical modeling to estimate standard errors in a survey sample, different proposed models used by other countries in similar surveys are summarized. The mathematical formulas for the different models are included in the next few paragraphs with a short explanatory discussion.

2.1 Indirect methods of estimating standard errors

The use of indirect methods to estimate standard errors has been practiced with satisfactory results by several countries, including the USA, Australia and Sweden. Different models are used according to suitability and preferences. Part of this project is to test and examine some of these models for suitability on typical South African data sets like the data sets made available by Statistics South Africa.

The data sets used in the research project are the October Household surveys (OHS) of Statistics South Africa of 1995, 1996 and 1997 and the Victims of Crime survey (VOC) of 1998. The OHS consists of more than one section, including the persons section, the workers section and the household section. The sample sizes for the OHS of 1995 and 1997 were 30 000 households and for the OHS of 1996 they were 16 000 households. The VOC reports on the crimes committed against members of the household, including the violent and non-violent crimes, in South Africa. The sample for the VOC consisted of 4000 households from which one person, aged 16 years or older, was selected to be interviewed. This person was chosen using a table of random numbers and fieldworkers were instructed to interview only this person.

2.2 Levels of domains of interest

Usually South African data sets, e.g. the workers subset of the OHS with a target population of all economically active people between the ages of 15 and 65 years, have a unique composition. This is due to the inclusion of four different race groups in the data sets, the different provinces being covered as well as the substantial differences between urban and rural areas in South Africa (see page 11). All these different classes lead to a large variety of domains of interest in SA data sets, in addition to the usual gender by age type of domains as shown in Table 2. This adds to the complexity of calculating standard errors.

Table 2: Table of the levels of domains used in this research project

Subclass	Number of categories	Type of class
RSA	1	Segregated class
Province	9	Segregated class
Urban / Rural (U/R)	2	Segregated class
E A type	5	Segregated or cross-class
Race	4	Mixed class
Gender	2	Cross-class
Age group	3	Cross-class
Province by U/R	18	Segregated class

Province by gender	18	Cross-class within segregated class
Province by age group	27	Cross-class within segregated class
Province by race	36	Mixed class within segregated class
U/R by race	8	Mixed class within segregated class
U/R by gender	4	Cross-class within segregated class
U/R by age group	6	Cross-class within segregated class
Race by gender	8	Cross-class within mixed class
Race by age group	12	Cross-class within mixed class
Gender by age group	6	Cross-class

A large number of categories for a subclass may have the result that the sample sizes of some of the subclass categories become too small to be included in the modeling procedure.

2.3 Models proposed by other countries

As already been mentioned, different models for the estimation of standard errors are used in practice by other countries. Some of these models are discussed in the next few paragraphs and will be fitted to the data in the next chapter, viz. Generalized Variance Functions used by the United States, models proposed by Lepkowski and models used by the Australian Bureau of Statistics.

2.3.1 The United States

In the USA, Generalized Variance Functions were used to estimate standard errors for the Scientists and Engineers Statistical Data System (SESTAT) survey which combines information from three National Science Foundation-sponsored surveys (Cox, Jang and Edson; 1993):

- The National Survey of College Graduates
- The Survey of Doctorate Recipients, and
- The National Survey of Recent College Graduates

Two other surveys in the United States that also make use of Generalized Variance Functions are the Current Population Survey (CPS) and the National Health Interview Survey (HIS) (Valliant; 1987).

Generalized Variance Functions (GVFs) are mathematical functions that describe the relationship between a population parameter (such as a population total) and its corresponding variance. GVFs provide users with a quick and simple way to model standard errors. The user inserts the estimated value of the statistic of interest into the fitted GVF model to generate a model-based approximation of the variance.

A GVF depends on the assumption that the relative variance of an estimated population parameter, $\hat{Y}$, is a decreasing function of the magnitude of the estimate:

$$\text{Rel}\,Var(\hat{Y}) = \alpha + \beta Y^{-1} \qquad (8)$$

where α and β are known as the GVF parameters (Johnson and King; 1987).

The relationship (8) can be derived as follows. Consider a sample of n units from a population of size N, where $\hat{P}$ denotes the estimate of the proportion $P = \frac{Y}{N}$ of a population characteristic, and Y is some counting variable measuring the occurrence of the characteristic, hence Y is the number of units in a certain class, c. Let D be the design effect accounting for departures from simple random sampling and $Var(\hat{P})$ the population variance of $\hat{P}$. The probability sampling **relative variance**[1] of $\hat{P}$ is then:

$$\begin{aligned} \text{Rel}Var(\hat{P}) &= \frac{Var(\hat{P})}{P^2} \\ &= \frac{DP(1-P)}{nP^2} \\ &= \frac{D(1-P)}{nP} \\ &= \frac{D - DP}{n\frac{Y}{N}} \\ &= \frac{-D}{n} + \frac{ND}{nY} \end{aligned} \qquad (9)$$

Formula (9) also holds for $\text{Rel}Var(\hat{Y})$ where $\hat{Y}$ is the estimated number of units in a certain class, c:

$$\begin{aligned} \text{Rel}Var(\hat{Y}) &= \text{Rel}Var(\hat{P}N) \\ &= \frac{Var(\hat{P}N)}{P^2N^2} \\ &= \frac{N^2Var(\hat{P})}{P^2N^2} \\ &= \frac{-D}{n} + \frac{ND}{nY} \end{aligned}$$

which is of the form $\text{Rel}Var(\hat{Y}) = \alpha + \beta Y^{-1}$ where $\alpha = \frac{-D}{n}$ and $\beta = \frac{ND}{n}$.

To derive the estimated standard error from this model is very simple (Valliant; 1987):

$$\text{var}(\hat{Y}) = rel\ \text{var}(\hat{Y}) \times \hat{Y}^2$$

where $rel\ \text{var}(\hat{Y})$ is the estimated relative variance of $\hat{Y}$.

$$\begin{aligned} \therefore \text{var}(\hat{Y}) &= \hat{\alpha}\hat{Y}^2 + \hat{\beta}\hat{Y} \\ \therefore se(\hat{Y}) &= \sqrt{\hat{\alpha}\hat{Y}^2 + \hat{\beta}\hat{Y}} \end{aligned} \qquad (10)$$

[1] **Definition of relative variance: RelVar = $(CV)^2$ where CV is the population coefficient of relative variation**

In the same way the estimated standard error of $\hat{P}$ can be derived (Valliant; 1987):

$$\begin{aligned}\text{var}(\hat{P}) &= rel\ \text{var}(\hat{P}) \times \hat{P}^2 \\ &= \frac{-\hat{P}^2 d}{n} + \frac{\hat{P}^2 Nd}{nY} \\ &= \frac{-\hat{P}^2 d}{n} + \frac{\hat{P}^2 d}{n\hat{P}} \\ &= \frac{\hat{P}d}{n}\left(1-\hat{P}\right) \\ \therefore se(\hat{P}) &= \sqrt{\frac{\hat{\beta}}{N}\hat{P}(1-\hat{P})}\end{aligned} \tag{11}$$

where $\hat{P}$ is the estimated proportion and $\hat{\beta} = \frac{Nd}{n}$ with d an estimate of D.

Obtaining the GVF parameters requires the calculation of a number of variances of the survey statistics through direct calculation methods, e.g. in the SESTAT survey the successive difference replication method was used. The α and β parameters are then estimated by fitting the model to these survey estimates and their variances.

2.3.2 Models proposed by Lepkowski

Lepkowski introduced the use of Generalized Variance Functions to estimate standard errors in a survey report. He proposed a model exactly the same as model (8) (Lepkowski; 1998).

Other models proposed by Lepkowski are mathematical derivations from model (8).

First derived model: From the proof of formula (9) follows

$$\begin{aligned}\text{Rel}Var(\hat{P}) &= \frac{D(1-P)}{nP} \\ &= \frac{DN(1-P)}{nY}\end{aligned}$$

and when P is small

$$\doteq \frac{DN}{nY}$$

This approximation gives a model of the form:

$$\text{Rel}Var(\hat{P}) = qY^{-1} \tag{12}$$

where $q = \frac{DN}{n}$

In the same way as on page 17 it can be shown that $\text{Rel}Var(\hat{P}) = \text{Rel}Var(\hat{Y}) = qY^{-1}$.

Further derived models:

Formula (12) can be converted into the coefficient of variance by taking the square root:

$$\text{RelVar}(\hat{P})^{\frac{1}{2}} = (qY^{-1})^{\frac{1}{2}}$$

$$\frac{1}{2}\log(\text{RelVar}(\hat{P})) = \frac{1}{2}\log(q) - \frac{1}{2}\log(Y)$$

$$\log(\text{RelVar}(\hat{P})) = q' + k\log(Y) \qquad (13)$$

where $q' = \log(q)$ and $k = -1$
(Ghangurde; 1981) and (Kalton; 1977)

Formula (13) is used by the Australian Bureau of Statistics and by Statistics Canada. (Valliant; 1987)

2.3.3 The Australian Bureau of Statistics (ABS)

The ABS has derived mathematical models by applying smoothing regression techniques on the standard errors calculated through split-half techniques (ABS; 1997).

The following assumptions were made in deriving the models: Simple random sampling without replacement (SRSWOR) is used to draw the sample of size n from the population of size N. Let Y_c denote the number of people in category c in the population (e.g. number of unemployed people) and be estimated by:

$$\hat{Y}_c = N\hat{P}_c \qquad (14)$$

where $\hat{P}_c$ is the estimated proportion of the sample in category c. (Note that $\hat{Y}_c$ is the sum of an indicator variable that takes on the value 1 if the sample unit is in category c and 0 otherwise.) Furthermore, the estimate $\hat{Y}_c$ is unbiased with an expected value of:

$$E(\hat{Y}_c) = NP_c \qquad (15)$$

$\hat{P}_c$ has a variance of:

$$Var(\hat{P}_c) = \frac{1}{n}\left(\frac{N-n}{N-1}\right)P_cQ_c$$

(Cochran; 1977)

Thus giving a variance for $\hat{Y}_c$ of:

$$Var(\hat{Y}_c) = \frac{N^2}{n}\left(\frac{N-n}{N-1}\right)P_cQ_c \qquad (16)$$

where $P_c = \frac{Y_c}{N}$ and $Q_c = 1 - P_c$

The relative standard error % $(RSE\%)$ of $\hat{Y}_c$ is:

$$
\begin{aligned}
RSE\%(\hat{Y}_c) &= \frac{\sqrt{Var(\hat{Y}_c)}}{Y_c}\times 100 \\
&= \frac{\sqrt{\frac{N^2}{n}\left(\frac{N-n}{N-1}\right)P_cQ_c}}{NP_c}\times 100 \\
&= \sqrt{\frac{\frac{N^2}{n}\left(\frac{N-n}{N-1}\right)P_cQ_c}{N^2P_c^{\,2}}}\times 100 \\
&= \sqrt{\frac{(N-n)Q_c}{n(N-1)P_c}}\times 100 \\
&\cong \sqrt{\frac{(N-n)Q_c}{nNP_c}}\times 100 \\
&= \sqrt{\frac{1-f}{f}\frac{Q_c}{Y_c}}\times 100
\end{aligned}
\qquad (17)
$$

where $f = \frac{n}{N}$ denotes the sampling fraction (ABS; 1997).

However, the survey sample is usually not drawn with SRSWOR and to compensate for the design effect, formula (17) should be adapted to take the design effect into account:

$$
RSE\%(\hat{Y}_c) \cong \sqrt{d}\sqrt{\frac{(1-f)}{f}\frac{Q_c}{Y_c}}\times 100 \qquad (18)
$$

where d is an estimate of D, the design effect.

In exactly the same manner one can derive the relative standard error % for the ratio estimator $\hat{R}$. ($\hat{R} = \frac{\hat{Y}_c}{\hat{X}}$ with $\hat{X}$ as the estimated number of people in a domain, e.g. $\hat{X}$ is the estimated number of economically active men, and $\hat{Y}_c$ as the estimated number of people in category c within the same domain, e.g. $\hat{Y}_c$ is the estimated number of unemployed men. Thus $\hat{Y}_c$ and $\hat{X}$ are sums of indicator variables, y_i and x_i respectively.)

The following variance-formula of $\hat{R}$ is used (Cochran; 1977):

$$
V(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2}\frac{1}{N-1}\sum_i^N (y_i - Rx_i)^2
$$

Since $\sum_i^N (y_i - Rx_i)^2 = \left[XR(1-R)^2 + X(1-R)(-R^2)\right] = XR(1-R)$ (which is equivalent to Cochran, chapter 3, for the Binomial distribution), the variance-formula of $\hat{R}$ is:

$$V(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2}\frac{XR(1-R)}{N-1} \tag{19}$$

Thus:

$$\begin{aligned} RSE\%(\hat{R}) &= \frac{\sqrt{Var(\hat{R})}}{R}\times 100 \\ &= \sqrt{\frac{1-f}{n\left(\frac{X}{N}\right)^2}\left(\frac{XR(1-R)}{R^2(N-1)}\right)}\times 100 \\ &\cong \sqrt{\frac{1-f}{n}\left(\frac{NXR(1-R)}{R^2X^2}\right)}\times 100 \qquad \text{Note } \frac{N}{N-1}\approx 1 \\ &= \sqrt{\frac{1-f}{f}\frac{(1-R)}{RX}}\times 100 \\ &= \sqrt{\frac{1-f}{f}\frac{(1-R)}{Y_c}}\times 100 \end{aligned}$$

Taking the design effect into account gives:

$$RSE\%(\hat{R}) \doteq \sqrt{d}\sqrt{\frac{1-f}{f}\frac{(1-R)}{Y_c}}\times 100 \tag{20}$$

If the natural logarithm is taken, we get:

from (18) $$\ln RSE\%(\hat{Y}_c) = a_c - \frac{1}{2}\ln(Y_c) + \frac{1}{2}\ln(1-P_c) \tag{21}$$

or from (20) $$\ln RSE\%(\hat{R}) = a'_c - \frac{1}{2}\ln(Y_c) + \frac{1}{2}\ln(1-R) \tag{22}$$

where the factors a_c and a'_c depend on the category considered through the design effect d and on the population size through f (ABS; 1997).

If a_c or a'_c is correlated with Y_c, P_c or R, the coefficients of $\ln(Y_c)$, $\ln(1-P_c)$ and $\ln(1-R)$ would be different from 0.5.

When the model is fitted to the data, population parameters are replaced by their estimated values. Changing from percentage to proportion:

Model 1

$$\ln(cv(\hat{Y}_c)) = a + b\ln(\hat{Y}_c) + c\ln(1-\hat{P}_c)$$

or

$$\ln(cv(\hat{R})) = a' + b'\ln(\hat{Y}_c) + c'\ln(1-\hat{R})$$

where cv denotes the estimated coefficient of relative variation.

$\hat{P}_c$ and $\hat{R}$ are, in addition to $\hat{Y}_c$, independent variables in the above models adding to the degree of difficulty when the models are used in practice. If the third terms are omitted, the above models are simplified to:

Model 2

$$\ln cv(\hat{Y}_c) = a + b\ln(\hat{Y}_c)$$

or

$$\ln cv(\hat{R}) = a' + b'\ln(\hat{Y}_c)$$

Whether it is justified to omit the third terms in Model 1 depends on the results when Model 1 and Model 2 are fitted to the data respectively and compared. If Model 2 produces poor results in comparison with the results from Model 1, it is advisable not to use Model 2.

Take note of the fact that the only difference between the standard error models for $\hat{Y}_c$ and $\hat{R}$, is that the model parameters a and b in the standard error model for $\hat{Y}_c$ differ from the model parameters a' and b' in the standard error model for $\hat{R}$. The implication is that even if the standard error estimate for $\hat{R}$ is required, the value of the estimated total of category c, $\hat{Y}_c$, which in this case refers to the value of the numerator in the estimated ratio $\hat{R}$, must be substituted into the standard error model, $\ln cv(\hat{R}) = a' + b'\ln(\hat{Y}_c)$.

In the cases where the value $\hat{P}_c$ (or $\hat{R}$) approaches 1, Model 2 tends to result in a larger value $cv(\hat{Y}_c)$ or $cv(\hat{R})$ than really exists, and this consequently gives rise to outliers (ABS; 1997). One possible solution to compensate in Model 2 for the additional term in Model 1, is to include a quadratic term into Model 2, leading to Model 3:

Model 3

$$\ln cv(\hat{Y}_c) = a + b\ln(\hat{Y}_c) + c\left(\ln(\hat{Y}_c)\right)^2$$

or

$$\ln cv(\hat{R}) = a' + b'\ln(\hat{Y}_c) + c'\left(\ln(\hat{Y}_c)\right)^2$$

using $\left(\ln(\hat{Y}_c)\right)^2$ as a rough substitute for $\ln(1-\hat{P}_c)$, or as a rough substitute for $\ln(1-\hat{R})$, depending on the population parameter ($\hat{Y}_c$ or $\hat{R}$) for which the standard error is being modeled (ABS; 1997). Again these models are the same except for the respective model parameters, a, b, c and a', b', c' that are different.

The question remains: Which of the above proposed models will give the best results when fitted to typical South African data sets, like the OHSs? The next section addresses this question.

3. The Modeling Procedure

The different models that were introduced in the previous chapter are now fitted to the data. The results of the different models considered are compared with each other in order to find the best suitable standard error model for the data considered. The modeling procedure shortly consists of the direct calculation of standard error estimates for a number of study variables in the survey, the fitting of the different models to the data by means of regression modeling, the identification of outliers in the data and the comparison of the modeling results.

3.1 Estimation of the model parameters

The estimation of the parameters in the standard error models considered in chapter 2 requires the calculation of the variance estimates of a number of typical survey estimates through direct methods in order to be included in the model fitting procedure. Although it is not necessary to calculate the variance of each survey estimate directly, a larger number of related survey estimates and their variances that cover a wide range of the domains of interest would contribute to a more representative model.

There are several different ways to calculate variance estimates directly. SESTAT made use of successive differences techniques and resampling methods such as random groups, balanced repeated replication and jackknife replication. The ABS used split-half techniques where the sample is split into two similar sections to calculate estimates of the standard errors directly.

In this research project SAS programs (refer to pages 11 to 14) based on the formulae of Verma (Verma; 1982) were used to calculate the estimated coefficients of relative variation and estimated standard errors of complex sample estimates for different domains of interest. By simply changing the categorical variable criteria in the program macro, it is very easy to calculate standard error estimates and coefficients of relative variation for every desired set of domains of interest of a specific estimate.

For a variety of population parameters or characteristics the standard error model is fitted to the estimated coefficients of relative variation obtained for the set of domains of interest, by making use of Least Squares (LS) regression. Survey estimates of both large values and small values should be included in the model fitting procedure. This will contribute to a good fit of the model at large, small and in-between values of the estimates.

3.2 Procedure of fitting data to the model

There are many software packages that make regression modeling very easy, e.g. STATISTICA, STATSGRAPHICS, SPSS, SAS, MICROSOFT EXCEL and many more. SAS INSIGHT was used to do model fitting for this project.

Choosing the best model mainly depends on finding the model with the highest coefficient of determination, R^2. The R^2-value gives the proportion of the variability in the dependent variable that can be explained by the fitted regression line. If the fitting results are not satisfactory, it can either be due to the existence of outliers or to a model that is not suitable for the data.

After the fitting procedure, the outliers must be identified, if there are any. If it is justified to exclude the outliers from the calculations, it is recommended to repeat the fitting procedure without the outlier-observations. Consequently, it is very important to first try to establish the reason for the outlier's occurrence. It was found that most outliers occur because of one of the following reasons:

a) The sample size of the domain on which the estimate is based is too small. This happens when there are only a small number of records in the specific domain or when all the records are concentrated in only a few of the PSU's. However, it seems that there does not exist a definite cut off point in the size of the domain that can be identified as too small.

   However, it was found to be usually the case that if the sample size of a specific domain of interest was as small as 10, it had to be discarded from the data set or else it produced outliers in the data. Estimated proportions, $\hat{P}_c$ or $\hat{R}$, that are found to be close to 0 or 1, for modeled coefficients of relative variation from the model $\ln(cv(\hat{Y}_c)) = a + b\ln(\hat{Y}_c)$, generally resulted in values that differ largely from the direct calculated values.

   Statistics SA does not publish estimates for too small sample sizes of the domains of interest. Thus, these estimates can be excluded from the modeling procedure of standard errors.

   Another possible solution for some of these cases would be to use Model 3 (page 22) instead of the above model. The factor $\ln(1-\hat{P}_c)$ becomes important when $\hat{P}_c \approx 1$ or $\hat{R} \approx 1$. To compensate for this, a quadratic term, $\left(\ln(\hat{Y}_c)\right)^2$, is included in Model 3 and serves as a rough substitute for $\ln(1-\hat{P}_c)$ or $\ln(1-\hat{R})$ (ABS; 1997).

b) Survey estimates with direct calculated coefficients of relative variation estimates larger than 0.1, i.e. $$cv(\hat{X}) = \sqrt{\frac{\text{var}(\hat{X})}{\hat{X}^2}} > 0.1$$ may result in outliers, but it depends on the whole data set. However, for ratio estimation the estimate can be biased to the extent that the estimate becomes misleading when $cv(\hat{X}) > 0.15$, with X denoting the variable in the denominator of the ratio. Such cases should thus be excluded in the modeling procedure.

c) Outliers were also observed for subclasses of the domain under consideration where $\hat{P}_c$ or $\hat{R} \approx 1$ for some subclasses and $\hat{P}_c$ or $\hat{R} \approx 0$ for other subclasses. Such cases are for example " water on site " and " toilet on site " which are applicable to almost all households in the formal urban area, but at the same time, are applicable to almost none of the households in the informal urban area. Again Model 3 can be used in these cases.

3.3 Goodness of the fit and identification of outliers

Apart from the R^2-value as an indication of how well the model fits the data, there are other guidelines and tests that help with deciding if the model is suitable. These tests are easy to perform with the help of a statistical software package such as SAS INSIGHT.

One possibility is to investigate the distribution of the residuals. If the model is suitable for the data, the residuals would follow, or very nearly follow, a normal distribution. A Normal probability plot is very useful in indicating gross departures from normality, which can either be because the data does not fit the model, or because of the presence of outliers.

Another practical guideline to follow is to plot the standardized residuals versus the predicted values. Nearly all the residuals should lie between the -2σ and $+2\sigma$ confidence bands. In a good fit, the residuals will be scattered randomly around the X-axis with the larger concentration near the X-axis. Residuals lying outside of the 2σ bands could indicate the presence of outliers.

Alternatively, the absolute values of the standardized residuals are considered. Values larger than 2 could indicate outliers while values larger than 3 should be regarded as outliers; i. e. $|e_i^*|>3$ where e_i^* denotes the standardized residual.

Another measure to use in the evaluation of the fitted models, is the *absolute relative difference* (ARD). This test is expressed as $\frac{|Modeled\ se\ -\ Direct\ calculated\ se|}{Direct\ calculated\ se}*100$. The average ARD for the set of direct and modeled standard errors (se) quantifies the average distance between them as a percentage of the actual standard error. A better fit is indicated by smaller ARD means (Bieler and Williams; 1990). Take note that the ARD mean is calculated as the average of the different ARD values obtained for a specific model over all the different domains of interest. Therefor, ARD means should be compared globally between the different models where a smaller ARD mean indicates a better fit. All these tests will be discussed further in an example in the next paragraph.

3.4 Finding the best suitable model for the data

For illustrative purposes, the results of the fitted regression model on the 1997 October Household Survey Workers data set, are summarized and discussed below. The results of all the other investigated surveys are included in Appendix - A.

The data set used for the regression modeling procedure in SAS INSIGHT to find a suitable standard error model for the 1997 OHS Worker data set is in Appendix - B (discussed on pages 10 to 14). In Appendix - B the results of the Workers data set of the OHS of 1997 are summarized, after the necessary SAS programs were run to estimate the standard error and coefficient of relative variation for the study variable of interest (number of unemployed, and the unemployment rate in this case).

From SAS INSIGHT the following graphical output (page 26) regarding the estimated population total of unemployed in SA, $\hat{Y}_c$, is obtained when the model, $\ln(cv(\hat{Y}_c)) = a + b\ln(\hat{Y}_c)$, is fitted to the data in Appendix - B.

First set of fitting results:

Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population total unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$. (Source: OHS 1997 - Workers)

Model: $\ln(cv(\hat{Y}_c)) = 2.5078 - 0.4312\ln(\hat{Y}_c)$

Fig 1:

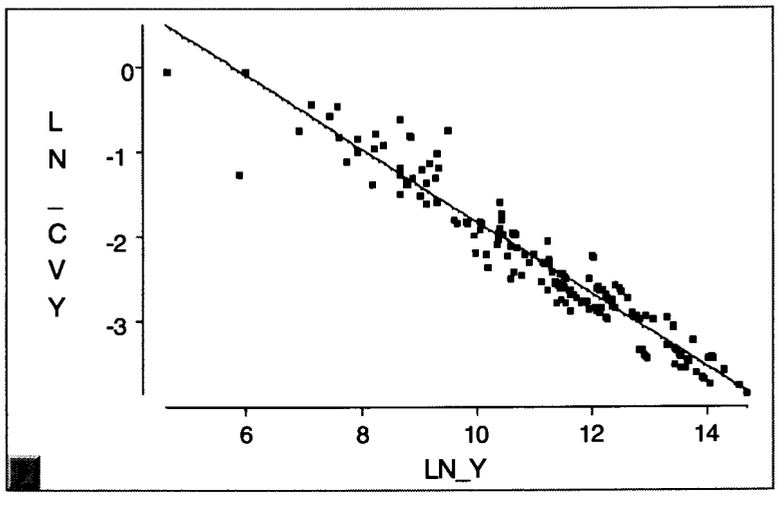


Table 3:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	133.8087	195	0.0529	0.9284	2529.1800	0.0001

Table 4:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5078	0.0974	25.7393	0.0001	.	0
LN_Y	1	-0.4312	0.0086	-50.2910	0.0001	1.0000	1.0000

Fig 2:

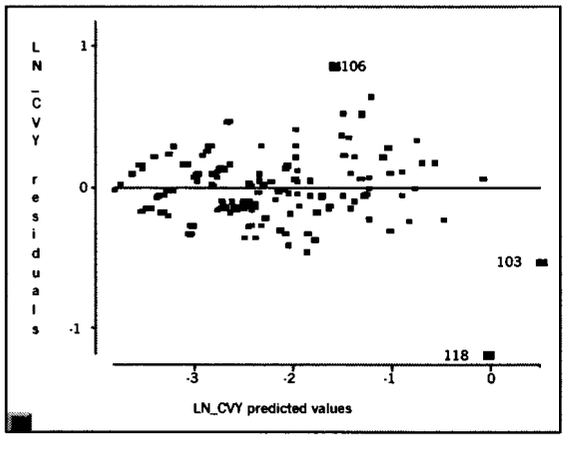
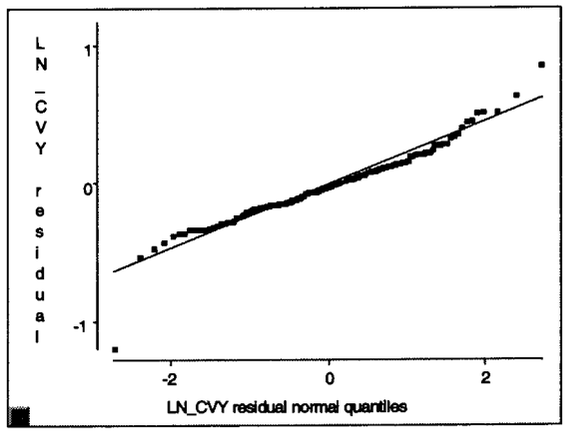


Fig 3:



A discussion of the results follows.

Fig 1: Plot of the linear relationship between $\ln(cv(\hat{Y}_c))$ as the dependent variable, and $\ln(\hat{Y}_c)$ as the predictor, where $\hat{Y}_c$ denotes the estimated population total of unemployed in South Africa.

In the model that is being fitted to the data: $\ln(cv(\hat{Y}_c)) = a + b\ln(\hat{Y}_c)$, a and b are the model parameters that should be estimated by using LS Regression. Figure 1 shows that a linear relationship between $\ln(cv(\hat{Y}_c))$ and $\ln(\hat{Y}_c)$ does exist. All the observations were included in this graph without excluding any outliers.

Table 3: The summary of the regression fit results

The R^2-value of 0.9284 shows that the model that was fitted on the data can explain almost 93% of the variation in $\ln(cv(\hat{Y}_c))$, giving evidence that the model is suitable for this data set.

Table 4: A summary of the estimated parameters

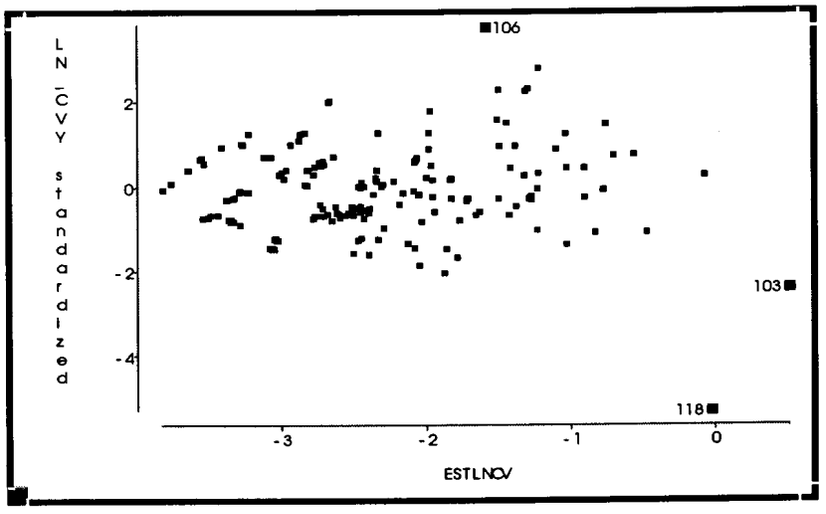
The small exceedance probabilities of 0.0001 for both parameters show that both the parameters are significant in the model.

Fig 2: Plot of the residual values of $\ln(cv(\hat{Y}_c))$ versus the predicted values of $\ln(cv(\hat{Y}_c))$.

The residual-versus-predicted values plot (Figure 2) serves as a test for outliers and to diagnose non-constant error variance. The residual values (take notice: not standardized residuals) seem to be randomly scattered around 0. There is a possibility that observations 103, 106 and 118 could be outliers, because they are lying outside the band containing the majority of residuals. These observations require further testing.

When the standardized residuals are plotted against the predicted values of $\ln(cv(\hat{Y}_c))$, observations 106 and 118 are lying substantially outside the 2σ bands (refer to Figure 2A).

Fig 2A: Standardized Residual Plot



The test $|e_i^*| > 3$ where e_i^* is the standardized residual, identifies values 106 (n=2) and 118 (n=1) as outliers. This could be because the subclass sample sizes in both cases are small. Value 103 (n=9) is not identified as an outlier. Therefor value 103 is not excluded from the data set while values 106 and 118 are, after which the regression modeling procedure is repeated.

Fig 3: Residual Normal Quantile Quantile plot of the residuals of $\ln(cv(\hat{Y}_c))$ versus the residual normal quantiles of $\ln(cv(\hat{Y}_c))$.

The Residual Normal QQ plot displays the extent to which the residuals are normally distributed. The empirical quantiles are plotted against the quantiles of a standard normal distribution. If the residuals follow a normal distribution, which is evident of a good fit, the points tend to fall along a straight line.

From Figure 3 it appears as if the residuals do follow a normal distribution with probable outlier observations at the upper – and the lower end of the plot. This gives further evidence that this model is suitable for this data set.

The next step is to repeat the whole fitting procedure, excluding the identified outliers, to see if there is a significant improvement in the fit.

On page 29 the second set of fitting results is given. The R^2-value increased to 0.9421. No other outliers could be identified. The obtained ARD mean for this model is 17.5%, which indicates a good fit when compared with much larger ARD means from other models (see Table 9). The model $\ln(cv(\hat{Y}_c)) = 2.588 - 0.4382\ln(\hat{Y}_c)$ can thus be accepted as a suitable model to approximate the standard errors for $\hat{Y}_c$, the estimated total number of unemployed people in a subclass for the workers data set of the 1997 OHS.

To derive the standard error from the model, the following conversion needs to be done:

$$\begin{aligned} cv(\hat{Y}_c) &= e^{2.588} \times e^{-0.4382\ln(\hat{Y}_c)} \\ &= e^{2.588} \times \left(\hat{Y}_c\right)^{-0.4382} \\ &= 13.30314 \times \left(\hat{Y}_c\right)^{-0.4382} \\ \therefore se(\hat{Y}_c) &= cv(\hat{Y}_c) \times \hat{Y}_c \\ &= 13.30314 \times \left(\hat{Y}_c\right)^{1-0.4382} \\ &= 13.30314\left(\hat{Y}_c\right)^{0.5618} \end{aligned}$$

If for example $\hat{Y}_c$ represents the estimated number of unemployed men in the Western Cape for 1997, we find from Appendix – B (page B-3) $\hat{Y}_c = 81091$. This value is substituted in the above formula:

$$\begin{aligned} se(\hat{Y}_c) &= 13.30314 \times (81091)^{0.5618} \\ &= 13.30314 \times 572.6125 \\ &= 7618 \end{aligned}$$

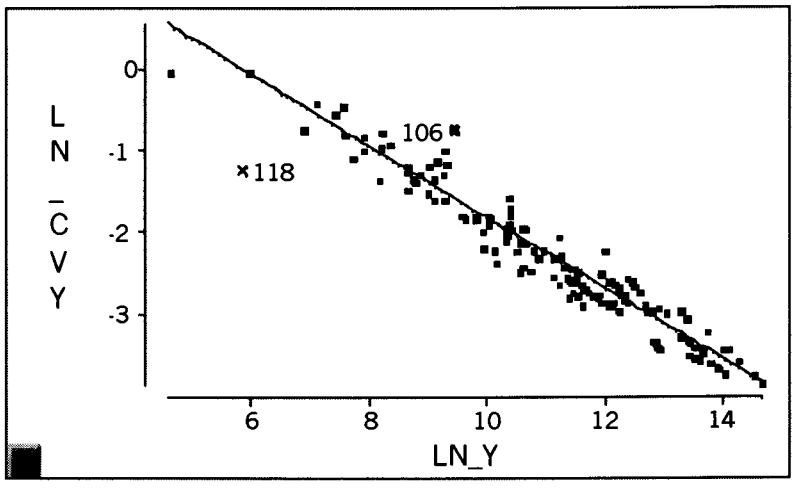
The direct calculated standard error estimate for the number of unemployed males in the Western Cape is 7394 (Appendix – B). The modeled standard error estimate of 7618, compares well with this value.

Second set of fitting results:

Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population total unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$. (Source: OHS 1997 - Workers)

Model: $\ln(cv(\hat{Y}_c)) = 2.588 - 0.4382\ln(\hat{Y}_c)$

Fig 4:



Observations 106 and 118 have been excluded from the calculations.

Table 5:

Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
	1	1	132.1242	193	0.0421	0.9421	3141.3556	0.0001

Table 6:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5880	0.0891	29.0534	0.0001	.	0
LN_Y	1	-0.4382	0.0078	-56.0478	0.0001	1.0000	1.0000

Fig 5:

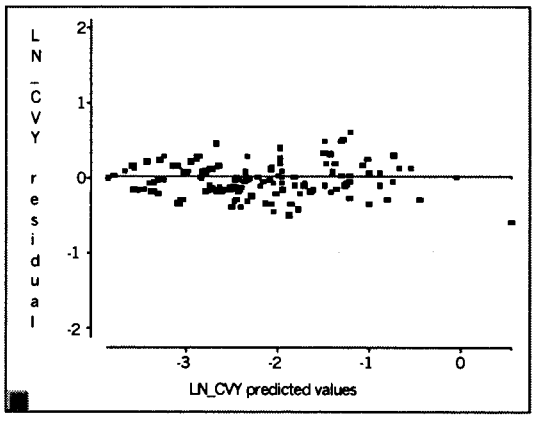
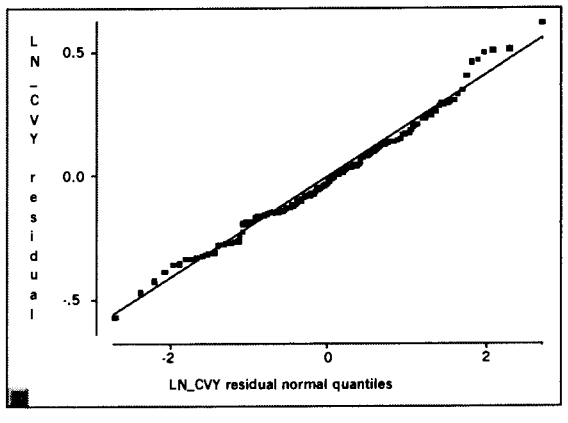


Fig 6:

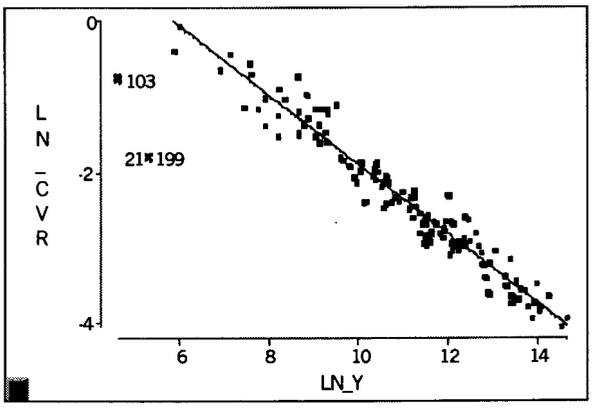


The same model fitting procedure can be done to find a standard error model for the ratio estimator, $\hat{R}$. The following results are obtained when the model derived for $\hat{R}$, $\ln(cv(\hat{R})) = a' + b' \ln(\hat{Y}_c)$, is fitted to the data with $\ln(cv(\hat{R}))$ as the dependent variable and $\ln(\hat{Y}_c)$ as the predictor in the model. $\hat{R}$ denotes the estimated unemployment rate in South Africa.

Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated number of unemployed according to the strict definition of unemployment. (Source: OHS 97 – Workers file)

Model: $\ln(cv(\hat{R})) = 2.7087 - 0.45851 \ln(\hat{Y}_c)$

Fig 7:



Observations 21, 103 and 199 are outliers and were excluded.

Table 7:

Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
	1	1	142.1090	194	0.0411	0.9469	3461.1108	0.0001

Table 8:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.7087	0.0897	30.5223	0.0001	.	0
LN_Y	1	-0.4585	0.0078	-58.8312	0.0001	1.0000	1.0000

Fig 8:

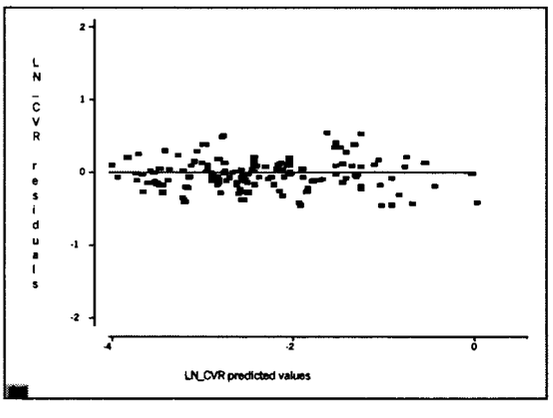


Fig 9:

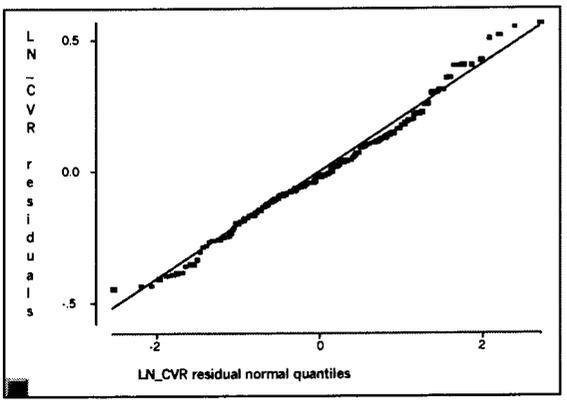


Fig 7, the plot of the linear relationship between $\ln(cv(\hat{R}))$ as the dependent variable, and $\ln(\hat{Y}_c)$ as the predictor, where $\hat{R}$ denotes the estimated unemployment rate for 1997 and $\hat{Y}_c$ denotes the estimated population total of unemployed for 1997, shows that a linear relationship between the natural logarithms of these estimates do exist. The R^2-value (in Table7) of approximately 0.95 is further evidence that the model is suitable for the data set (OHS 1997 – Workers data set).

Fig 8, the plot of the residual values of $\ln(cv(\hat{R}))$ versus the predicted values of $\ln(cv(\hat{R}))$, indicates that after observations 21, 103 and 199, which were identified as outliers, have been excluded from the data, no other outliers could be identified. An estimated ratio of approximately 0 exists for each of observations 21, 103 and 199, resulting in outliers as explained in reason c on page 24.

In Fig 9, the Residual Normal Quantile Quantile plot of the residuals of $\ln(cv(\hat{R}))$ versus the residual normal quantiles of $\ln(cv(\hat{R}))$, the plotted points tend to fall along a straight line, which confirms a good fit. Thus, $\ln(cv(\hat{R})) = 2.7087 - 0.45851\ln(\hat{Y}_c)$ is accepted as a suitable model to estimate standard errors for $\hat{R}$, the estimated unemployment rate for 1997.

Again this model must be converted to approximate the required standard error:

$$\begin{aligned} cv(\hat{R}) &= e^{2.7087} \times e^{-0.45851\ln(\hat{Y}_c)} \\ &= e^{2.7087} \times \left(\hat{Y}_c\right)^{-0.45851} \\ &= 15.0098 \times \left(\hat{Y}_c\right)^{-0.45851} \end{aligned}$$

$$\begin{aligned} \therefore se(\hat{R}) &= cv(\hat{R}) \times \hat{R} \\ &= 15.0098 \times \left(\hat{Y}_c\right)^{-0.45851} \times \hat{R} \end{aligned}$$

If for example $\hat{R}$ represents the estimated unemployment rate of men in the Western Cape for 1997 and $\hat{Y}_c$ the number of unemployed men in the same province for 1997, then we obtain from Appendix – B: $\hat{R}$ = 0.08906 and $\hat{Y}_c$ = 81091. These values are substituted in the above formula:

$$\begin{aligned} se(\hat{R}) &= 15.0098 \times (81091)^{-0.45851} \times 0.08906 \\ &= 15.0098 \times 0.0056 \times 0.08906 \\ &= 0.0075 \end{aligned}$$

The value, 0.0075, compares well to the standard error estimate of 0.0079 that was directly calculated with the SAS program for this study variable, $\hat{R}$, in the same subclass (unemployed men in the Western Cape) (see page B-3, Appendix – B).

3.5 Comparing the results of the different models fitted

After the modeling procedure for each different model discussed in the second chapter has been completed, the obtained R^2-values and ARD means are compared in order to find the model that gives the best results for the data sets considered. The results for the different models are shortly summarized.

Table 9: Table of the R^2-values* and ARD means when different models for $\hat{Y}_c$ (considering different study variables) were fitted to the data sets.

Equation number of formula used in fitting procedure		Models				
		US Formula [8]	**Lepkowski** Formula [12]	**ABS 1** Model 1	**ABS 2** Model 2	**ABS 3** Model 3
Study variable: Estimated number of unemployed (Strict definition of unemployment)						
Data set: OHS 1997 Worker data	**Fitted Model Parameters**	α = 0.0261 β =70.4501	q =75.6074	a =2.0713 b = -0.4094 c = -0.1035	a =2.588 b = -0.4382	a =1.6474 b = -0.2636 c = -0.0079
	R^2-value	0.3	0.43	0.93	0.94	0.93
	ARD Mean	99%	60%	17.7%	17.5%	17.8%
Study variable: Estimated number of unemployed (Strict definition of unemployment)						
Data set: OHS 1996 Worker data	**Fitted Model Parameters**	α = 0.0447 β =224.099	q =242.653	a =2.224 b = -0.3688 c =0.0134	a =2.1526 b =-0.364	a =1.4986 b = -0.2319 c = -0.0065
	R^2-value	0.52	0.6	0.93	0.93	0.93
	ARD Mean	93%	63%	16.2%	16.2%	15.6%
Study variable: Estimated number of unemployed (Strict definition of unemployment)						
Data set: OHS 1995 Worker data	**Fitted Model Parameters**	α = 0.0007 β =28.6229	q =29.7412	a =3.5271 b = -0.482 c =0.6854	a =1.1842 b =-0.3895	a =1.7685 b =0.2183 c = -0.0198
	R^2-value	0.56	0.6	0.90	0.89	0.78
	ARD Mean	81%	79%	29.3%	30%	70%
Study variable: Estimated number of households with different main lighting sources						
Data set: OHS 1997 Household data	**Fitted Model Parameters**	α = 0.019 β =414.856	q =427.428	a =3.4918 b = -0.4839 c =0.0922	a =2.5152 b = -0.4202	a =1.8187 b = -0.279 c = -0.0066
	R^2-value	0.45	0.89	0.97	0.97	0.96
	ARD Mean	200%	30.7%	17%	19.3%	19.1%

* Note that the R^2-value referred to as the squared multiple correlation coefficient is included. Outliers, where identified, were excluded from the data sets during the modeling procedure.

Study variable: Estimated number of households with different main lighting sources						
Data set: OHS 1996 Household data	**Fitted Model Parameters**	α = 0.0539 β =621.691	q =670.226	a =3.9174 b = -0.4883 c =0.1176	a =2.6678 b = -0.4072	a =1.6129 b = -0.1902 c = -0.0102
	R^2-value	0.75	0.82	0.97	0.97	0.96
	ARD Mean	59%	34.4%	13.3%	18%	17.7%
Study variable: Estimated number of households with different main lighting sources						
Data set: OHS 1995 Household data	**Fitted Model Parameters**	α = 0.0615 β =226.194	q =256.761	a =1.9094 b = -0.3971 c = -0.0448	a =2.5066 b = -0.4380	a =1.4386 b = -0.2111 c = -0.011
	R^2-value	0.4	0.79	0.97	0.97	0.97
	ARD Mean	200%	28.3%	16.8%	17.4%	16.5%
Study variable: Estimated number of households with different dwelling types						
Data set: OHS 1997 Household data	**Fitted Model Parameters**	α = 0.128 β =171.673	q =209.604	a =2.906 b = -0.4353 c =0.0252	a =2.4389 b = -0.3955	a = -1.0612 b = 0.3722 c = -0.0374
	R^2-value	0.45	0.52	0.86	0.89	0.89
	ARD Mean	235%	97%	32%	31%	28%
Study variable: Estimated number of households with different dwelling types						
Data set: OHS 1996 Household data	**Fitted Model Parameters**	α = 0.0907 β =337.309	q =397.826	a =2.9301 b = -0.4146 c =0.0115	a =2.7282 b = -0.3957	a = -1.0612 b = 0.3722 c = -0.0374
	R^2-value	0.6	0.65	0.88	0.90	0.92
	ARD Mean	51.9%	52%	28.6%	28.4%	26%
Study variable: Estimated number of crimes against households						
Data set: VOC 1998	**Fitted Model Parameters**	α = 0.155 β =4373.5	q =5187.73	a =2.5291 b = -0.3698 c = -0.1293	a =3.833 b = -0.4436	a = -1.1447 b = 0.3560 c = -0.0319
	R^2-value	0.82	0.84	0.92	0.91	0.89
	ARD Mean	32.6%	26%	12.7%	13.4%	20%
Study variable: Estimated number of crimes against persons						
Data set: VOC 1998	**Fitted Model Parameters**	α = 0.0091 β =6203.61	q =7290.38	a =1.9269 b = -0.3433 c = -0.1963	a =4.5238 b = -0.498	a = -8.3844 b = 1.4859 c = -0.0756
	R^2-value	0.74	0.86	0.96	0.95	0.97
	ARD Mean	26.4%	15.8%	10.4%	11.3%	10.3%

Comparing the R^2-values and ARD means in the above table, the ABS Model 1, 2 and 3 produce better results than the US and Lepkowski models (formula (8) and (12)). The ABS models seem to give better results in all the cases. The models of the US and Lepkowski produce low R^2-values and relative high ARD means in most cases. This leads to the conclusion that these models do not seem to fit the data well, especially when compared to the high R^2-values and much lower ARD means obtained from the ABS models.

On the other hand, the three different ABS models produce similar results, with comparable R^2-values and ARD means in most cases. It is not possible to isolate one model that performs best in each case in terms of the highest R^2-value and the lowest ARD mean. In the cases where a model produces the highest R^2-value, it does not necessarily have the lowest ARD mean and vice versa.

However, from a practical point of view, Model 2 deserves more credit for its simplicity (refer to page 22). For this reason, Model 2 is recommended to use in practice for the estimation of standard errors in the data sets considered.

The following table summarizes the results when different models for $\hat{R}$, the estimated ratio, are fitted to the data sets considered.

(Take note that although formula (9) and (12) were derived for the estimated proportion, $\hat{P}=\frac{\hat{Y}}{N}$, these formulas can also be used for $\hat{R}=\frac{\hat{Y}}{\hat{X}}$, where $\hat{X}$ is the estimated number of units in a domain in the population. The implication is that the model parameters for $\hat{R}$ in formula (9) and (12) differ from the model parameters for $\hat{P}$.)

Table 10: Table of the R^2-values* and ARD means when different models for $\hat{R}$ (considering different study variables) were fitted to the data sets.

Equation number of formula used in fitting procedure		Models				
		US Formula [9]	**Lepkowski** Formula [12]	**ABS 1** Model 1	**ABS 2** Model 2	**ABS 3** Model 3
Study variable: Estimated unemployment rate (Strict definition)						
Data set: OHS 1997 Worker data	**Fitted Model Parameters**	α = 0.0274 β =33.99	q =41.3597	a' =1.6156 b' = -0.3866 c' = -1.412	a' =2.7087 b' = -0.4585	a' = -0.4893 b' = 0.1231 c' = -0.0258
	R^2-value	0.1	0.16	0.90	0.95	0.91
	ARD Mean	97%	96%	27%	18%	20%

* Note that the R^2-value referred to as the squared multiple correlation coefficient is included. Outliers, where identified, were excluded from the data sets during the modeling procedure.

Study variable: Estimated unemployment rate (Strict definition)						
Data set: OHS 1996 Worker data	**Fitted Model Parameters**	α = 0.0333 β =245.417	q =259.247	a' =1.174 b' = -0.326 c' = -0.1954	a' =2.2642 b' = -0.3879	a' = 1.3437 b' = -0.2124 c' = -0.0082
	R^2-value	0.63	0.66	0.91	0.93	0.89
	ARD Mean	70%	69%	19%	20%	21%
Study variable: Estimated unemployment rate (Strict definition)						
Data set: OHS 1995 Worker data	**Fitted Model Parameters**	α = 0.0059 β =235.1	q =244.905	a' =3.2148 b' = -0.4676 c' = 0.1898	a' =2.1541 b' = -0.4099	a' = 0.967 b' = -0.1856 c' = -0.0103
	R^2-value	0.75	0.77	0.94	0.93	0.92
	ARD Mean	54%	51%	13%	14%	16%
Study variable: Estimated ratio of households with different main lighting sources						
Data set: OHS 1997 Household data	**Fitted Model Parameters**	α = 0.0085 β =411.35	q =417.28	a' =-0.1225 b' = -0.3018 c' = -0.295	a' =2.772 b' = -0.4642	a' = 4.073 b' = -0.7069 c' = 0.009
	R^2-value	0.87	0.90	0.94	0.93	0.80
	ARD Mean	34%	27%	24%	21%	39%
Study variable: Estimated ratio of households with different main lighting sources						
Data set: OHS 1996 Household data	**Fitted Model Parameters**	α = 0.444 β =574.92	q =615.028	a' =1.2702 b' = -0.371 c' = -0.2016	a' =2.9872 b' = -0.4563	a' = 3.629 b' = -0.5614 c' = 0.025
	R^2-value	0.75	0.81	0.82	0.92	0.80
	ARD Mean	56%	38%	37%	20%	40%
Study variable: Estimated ratio of households with different main lighting sources						
Data set: OHS 1995 Household data	**Fitted Model Parameters**	α = 0.0611 β =226.741	q =257.096	a' =1.2124 b' = -0.3917 c' = -0.1439	a' =2.8218 b' = -0.4806	a' = 1.2497 b' = -0.1234 c' = -0.0193
	R^2-value	0.68	0.78	0.91	0.96	0.91
	ARD Mean	69%	49%	21%	18%	23%
Study variable: Estimated ratio of crimes against households						
Data set: VOC 1998	**Fitted Model Parameters**	α = 0.0149 β =4130.18	q =4912.85	a' =1.7711 b' = -0.3386 c' = -0.2397	a' =4.0889 b' = -0.4669	a' = -2.3859 b' = 0.5849 c' = -0.0425
	R^2-value	0.71	0.79	0.86	0.87	0.84
	ARD Mean	42%	50%	19%	20%	27%

Study variable: Estimated ratio of crimes against persons						
Data set: VOC 1998	**Fitted Model Parameters**	α = 0.0082 β =6264.25	q =7239.77	a' =1.8857 b' = -0.3477 c' = -0.2195	a' =4.7898 b' = -0.5207	a' = -8.9856 b' = 1.5964 c' = -0.0807
	R^2-value	0.70	0.82	0.97	0.95	0.97
	ARD Mean	40%	25%	19%	19%	21%

Similar results as for $\hat{Y}$ are obtained when fitting the different models for $\hat{R}$. This can clearly be seen by comparing the results in Table 9 and Table 10. Model 2 can thus be used successfully in practice for the estimation of standard errors in the data sets considered.

Considering the results as listed in Table 9 and Table 10, a summary of conclusion follows.

Model proposed by the United States:

$$\text{Rel}Var(\hat{Y}) = \alpha + \beta Y^{-1}$$

$$\text{Rel}Var(\hat{R}) = \alpha' + \beta' Y^{-1}$$

(refer to formula (8) and (9))

The results obtained when this model was fitted to the data sets (OHS Worker data sets and Household data sets) were very disappointing. In most cases the fitted model resulted in a very unsatisfactory R^2-value of less than 0.6 and a very high ARD mean. The conclusion is that the GVF used by SESTAT and other US institutes is not suitable for these data sets.

Formula (12) (a mathematical derivation of the models proposed by **Lepkowski**):

$$\text{Rel}Var(\hat{Y}) = qY^{-1}$$

$$\text{Rel}Var(\hat{R}) = q'' Y^{-1}$$

The results when this model was fitted to the data sets were also disappointing. Although in some cases this model produces slightly better results than formula (8), the R^2-values are unsatisfactory low and the ARD means high (refer to Tables 9 & 10). The conclusion is that formula (12) is not the best suitable model for the data sets considered.

Model proposed by the Australian Bureau of Statistics:

$$\ln(cv(\hat{Y}_c)) = a + b\ln(\hat{Y}_c)$$

$$\ln(cv(\hat{R})) = a' + b'\ln(\hat{Y}_c)$$

where $\hat{Y}_c$ is the estimated number of people in category c and $\hat{R}$ is the estimated ratio of the people in the same category (refer to Model 2 on page 22).

Model 2 has proven to be one of the best models when fitted to the various data sets. Giving an R^2-value of not less than 0.9 in most cases and a relatively low ARD mean, it is safe to accept that this model is suitable for the data sets considered in this study.

Not denying the fact that Model 1 and Model 3 of the ABS also produce very good results, Model 2 has the advantage of the simplest formula, which makes Model 2 more user friendly. It seems that the contribution of the additional term in Models 1 and 3 towards better results is minimal when compared to the results from Model 2. For this reason Model 2 is chosen to estimate standard errors for the data sets considered.

Because the models for $\hat{Y}_c$ and $\hat{R}$ differ only in their respective model parameters, Model 2 has added practical value in the sense that both models for $\hat{Y}_c$ and $\hat{R}$ can be displayed on the same graph with the estimated coefficient of relative variation as the y-axis and the estimated total, $\hat{Y}_c$, as x-axis. This helps to make the presentation of these standard error models in the survey report more practical.

It is also found that, although the ABS had derived Model 2 for estimates of "person counts", the model performs equally well when estimates of counting variables in general are considered, e.g. counted households in the Household data set of the OHS or counted incidents of crimes in the VOC. Note again that $\hat{Y}_c$ in the derivation discussed on pages 19 to 21, is a counting variable: $y_{ci} = 1$ if the unit is in category c and $y_{ci} = 0$ otherwise. Consequently Model 2 gives satisfactory results in all the cases where a total or a ratio is estimated.

Another very important conclusion reached is that Model 2 seems to fit the data equally well for cross-classes, mixed classes and segregated classes. This result makes it possible to find one suitable standard error model for a study variable over all the domains of interest.

3.6 Illustration of results

The functionality of ABS Model 2 is investigated by means of a comparison between the directly calculated standard error estimates obtained from the previously mentioned SAS programs and the modeled standard error estimates by using Model 2. A number of the survey estimates of the 1997 OHS Workers data set and Household data set are displayed in the following tables for different domains of interest together with their directly calculated standard error estimates and modeled standard error estimates.

The following table (Table 11) illustrates the modeled standard error estimates for different domains by using Model 2 for the estimated number of unemployed in SA from the 1997 OHS Workers data set. As derived on pages 26 to 29, the fitted model is $\ln(cv(\hat{Y}_c)) = 2.588 - 0.4382 \ln(\hat{Y}_c)$. ARD values were also included in the table to illustrate the goodness of fit.

Table 11: Illustration of results

Study variable: Number of Unemployed –1997	Estimated value $\hat{Y}_c$	Fitted model: $\ln(cv(\hat{Y}_c)) = 2.588 - 0.4382\ln(\hat{Y}_c)$ $\therefore se(\hat{Y}_c) = e^{(\ln(cv(\hat{Y}_c)))} \times \hat{Y}_c$		
Domain		Directly calculated Standard errors	Modeled Standard errors	ARD (per domain)
African	2088753	49835	47257	5%
Coloured	209235	14749	12974	12%
Indian / Asian	41944	6055	5260	13%
White	77277	7824	7414	5%
Western Cape	185061	13824	12110	12%
Eastern Cape	303402	20422	15986	22%
Northern Cape	47209	4161	5621	35%
Free State	156583	9327	11025	18%
Kwazulu / Natal	474734	24780	20558	17%
North West	190619	11402	12313	8%
Gauteng	670552	32244	24960	23%
Mpumalanga	178189	10315	11855	15%
Northern Province	210861	11186	13031	17%

Table 12 contains the modeled standard error estimates for different domains by using the fitted model for the estimated unemployment rate in SA from the 1997 OHS Workers data set as derived on pages 30 to 31.

Table 12: Illustration of results

Study variable: Unemployment rate 1997	Estimated values		Fitted model: $\ln(cv(\hat{R})) = 2.7087 - 0.45851\ln(\hat{Y}_c)$ $\therefore se(\hat{R}) = e^{(\ln(cv(\hat{Y}_c)))} \times \hat{R}$		
Domain	$\hat{R}$	$\hat{Y}_c$	Directly calculated Standard errors	Modeled Standard errors	ARD (per domain)
African	0.2810	2088753	0.0051	0.0053	4%
Coloured	0.1525	209235	0.0083	0.0083	0%
Indian / Asian	0.0989	41944	0.0129	0.0113	12%
White	0.0406	77277	0.0040	0.0035	12.5%
Western Cape	0.1182	185061	0.0084	0.0068	19%
Eastern Cape	0.2907	303402	0.0182	0.0134	26%

Northern Cape	0.1854	47209	0.0173	0.0200	16%
Free State	0.2035	156583	0.0129	0.0127	2%
Kwazulu / Natal	0.2282	474734	0.0111	0.0086	23%
North West	0.2407	190619	0.0128	0.0137	7%
Gauteng	0.2168	670552	0.0095	0.0069	27%
Mpumalanga	0.2445	178189	0.0134	0.0144	8%
Northern Province	0.2626	210861	0.0138	0.0143	4%

From the 1997 OHS Household file, the estimated values for **dwelling-type = "formal house or brick structure on separate yard or stand" according to province** (Table 13) and **main water source = "piped (tap) water in dwelling" according to province** (Table 14), were considered. (Regression modeling details for these study variables are given in Appendix-A, pages A-7 and A-9.) Directly calculated standard error estimates and modeled standard error estimates are listed.

Table 13: Illustration of results

Study variable: **Dwelling Type** = Formal house or brick structure on separate yard or stand	Estimated value $\hat{Y}_c$	Fitted model: $\ln(cv(\hat{Y}_c)) = 2.4389 - 0.3955\ln(\hat{Y}_c)$ $\therefore se(\hat{Y}_c) = e^{(\ln(cv(\hat{Y}_c)))} \times \hat{Y}_c$		
Domain		Directly calculated Standard errors	Modeled Standard errors	ARD (per domain)
Western Cape	633402	25698	36840	43%
Eastern Cape	685917	26561	38657	46%
Northern Cape	155996	4804	15792	229%
Free State	391459	17636	27541	56%
Kwazulu / Natal	912224	35148	45928	31%
North West	553899	15417	33971	120%
Gauteng	1321969	38177	57475	51%
Mpumalanga	429799	13637	29141	113%
Northern Province	733840	17946	40268	124%

Table 14: Illustration of results

Study variable: **Main Water source** = Piped (tap) water, in dwelling	Estimated value $\hat{Y}_c$	Fitted model: $\ln(cv(\hat{Y}_c)) = 2.2365 - 0.3889\ln(\hat{Y}_c)$ $\therefore se(\hat{Y}_c) = e^{(\ln(cv(\hat{Y}_c)))} \times \hat{Y}_c$		
Domain		Directly calculated Standard errors	Modeled Standard errors	ARD (per domain)
Western Cape	776426	23638	41665	76%
Eastern Cape	336955	26641	25155	6%
Northern Cape	95346	7098	11727	65%
Free State	247448	20943	20872	1%
Kwazulu / Natal	640290	38002	37081	3%
North West	188721	18232	17719	3%
Gauteng	1243752	43696	55395	27%
Mpumalanga	221891	19003	19541	3%
Northern Province	114416	19275	13094	32%

Comparing the modeled standard error estimates with the direct calculated standard error estimates as illustrated in the above tables leads to a number of conclusions. In most cases, the different standard error estimates (modeled and directly calculated) compare very well. In many cases the ARD values (per domain) are below 30% and in some cases even below 10% indicating a good fit. However, it should be stressed that the ARD means give a global indication of the goodness of the fit and should be used to compare different models with each other rather than using individual ARD values per domain.

In the cases where direct and modeled standard error estimates do not compare well (e.g. in Table 13 for the study variable "dwelling Type = Formal house or brick structure on separate yard or stand" and in Table 14 for the study variable "Main Water source = Piped (tap) water"), they are usually of the same magnitude. However, the Northern Cape with an ARD value of approximately 229%, has a very large difference between its direct and modeled standard error estimates. Reason c for the occurrence of outliers as explained on page 24, provides an explanation why the results in Table 13 and 14 may seem less evident of a good fit. In Table 13 the situation exists where almost all the households with formal houses are found in the urban areas and none in the rural areas. This is exactly the situation explained in reason c. E.g. the total sample size in the Northern Cape is 1459 and 1229 of these households are in the category " Formal house or brick structure", giving an estimated ratio of 0.84. The implication is that only 0.16 of the households in the Northern Cape have other dwelling-types, resulting in estimated ratios of approximately 0. Nevertheless, an R^2-value of 0.89 and an ARD mean of 31% were obtained (refer to Table 9). Compared to the R^2-values and ARD means given by the other models considered for this study variable, Model 2 seems to fit the data well and produces acceptable results. The same explanation applies to the results in Table 14.

4. Presentation Methods

There are numerous ways to present standard errors in a survey report. The main requirements for the successful presentation of standard errors in a report are: the method should be cost effective in the sense of taking up as few pages as possible in the publication, easy to apply for the statistician and simple enough for the users to understand and use.

A short introduction to some of the methods adopted by other countries is given along with an example of each. A few of the advantages and disadvantages of each specific presentation method are also discussed. The different presentation methods that are considered are: a table with estimated model parameter values, a table with the estimated standard errors according to the size of the estimates, a table with estimated coefficients of relative variation and factor-lines, and formulas and graphs.

4.1 A table with estimated parameter values

The U.S. Bureau of the Census used the following method in the 1997 National Survey of College Graduates (Cox, Jang and Edson; 1993).

Having fitted a suitable model to approximate standard errors, the resulting estimated model parameters are displayed in a parameter table in the publication. Each new study variable in the survey, with its own model parameters, becomes an entry in the table.

The following steps describe the procedure to determine the standard error estimates of an estimated total or percentage (Finamore; 1999):

- Substitute the estimated total or percentage ($\hat{Y}_c$ or $\hat{R}$) into the standard error model that is provided;
- Find the table entry for the study variable of interest. If different models were fitted according to domains of interest, make sure to use the appropriate model parameters for the subclass on which the estimate is based;
- Substitute the parameter estimates into the model;
- Compute the approximate standard error.

The following example demonstrates the use of the parameter table for calculating the standard error of the estimate of the number of unemployed men in the Western Cape, in the 1997 OHS Worker data set.

4.1.1 Example:

Table 15: Parameter Table for the worker data set and the household data set of the October Household Survey of 1997.

Study Variables According to different domains	Model Coefficients for $\hat{Y}_c$: $\ln(cv(\hat{Y}_c)) = a + b\ln(\hat{Y}_c)$		Model Coefficients for $\hat{R}$: $\ln(cv(\hat{R})) = a' + b'\ln(\hat{Y}_c)$	
	Intercept a	Slope b	Intercept a'	Slope b'
Unemployed Strict definition	2.5880	-0.4382	2.7087	-0.4585
Unemployed Expanded def.	2.8358	-0.4623	2.6269	-0.4601
Dwelling Type	2.4389	-0.3955	2.7167	-0.4297
Water Source	2.2365	-0.3889	2.3443	-0.4067
Light Source	2.5152	-0.4202	2.772	-0.4642

Model: $\ln(cv(\hat{Y}_c)) = a + b\ln(\hat{Y}_c)$ and $\ln(cv(\hat{R})) = a' + b'\ln(\hat{Y}_c)$

where $\hat{Y}_c$ denotes the estimated total and $\hat{R}$ denotes the estimated ratio.

The above models were fitted on the data for $\hat{Y}_c$ and $\hat{R}$ respectively.

To estimate the standard error for the **estimated total of unemployed men in the Western Cape,** proceed as follows:

Obtain the estimate of the total number of unemployed males in the Western Cape for 1997 according to the strict definition of unemployment (Appendix – B, page B-3):

$$\hat{Y}_c = 81091$$

(Take note that this estimated value is part of the preliminary results and differs from the final released results by StatsSA due to weighting differences as mentioned on page 3.)

From the above table the parameter values are:

$$\hat{a} = 2.588 \text{ and } \hat{b} = -0.4382$$

Now we have the model:

$$\ln(cv(\hat{Y}_c)) = 2.588 - 0.4382\ln(81091)$$
$$\ln(cv(\hat{Y}_c)) = -2.3651$$

The standard error can be calculated with the following conversion:

$$se(\hat{Y}_c) = e^{(\ln(cv(\hat{Y}_c)))} \times \hat{Y}_c$$
$$se(\hat{Y}_c) = 7618$$

The direct calculated standard error for this estimate which is based on the subclass: gender = male and province = Western Cape is 7394 which compares well to the modeled standard error.

The standard error for $\hat{R}$ can be calculated in the same way.

Advantages:

- The method is fairly easy for the statistician to apply and is easy to understand.
- The possibility of including a separate pair of parameters for each new domain of interest may contribute to a higher level of accuracy in the modeling of standard errors.

Disadvantages:

- The more study variables and the more domain possibilities there are, the more parameter sets must be included in the table. This takes up space in the publication and can be time consuming. It also complicates the readability of the table.
- The method requires that the user is familiar with the substitution of the correct pair of parameters into the model and the calculation of the standard error with the formulas provided. Thus, this method requires a considerable amount of work to obtain the standard error required.

4.2 A table with the standard errors according to the size of the estimate

A table that consists of the estimated standard errors according to size of the survey estimates and the confidence intervals of a specific level of significance, is published. These standard errors can be estimated with the suitable model or calculated directly if the size of the data set in terms of number of study variables allows it.

In the example the fitted model was used to estimate the published standard errors and the associated confidence intervals were calculated on a 95% level of significance.

From the table, the user is expected to find the estimate that is nearest in size to the estimate from the survey whose standard error is desired. Note that in the table it is the estimate that is just larger in size that should be chosen rather than the one just smaller than the estimate of interest. The conservative approach should be followed whenever standard errors are concerned. However, the chosen estimate must compare realistically with the survey estimate.

4.2.1 Example:

A table with standard errors for the worker data set of the 1997 OHS is constructed according to the strict definition of unemployment. The standard errors are calculated using the fitted model:

$$\ln\left(cv(\hat{Y}_c)\right) = 2.588 - 0.4382\ln(\hat{Y}_c)$$

and the conversion formula:

$$se(\hat{Y}_c) = \exp\left(\ln(cv(\hat{Y}_c))\right) \times \hat{Y}_c$$

The confidence intervals are then calculated with the following formula at a level of 95% significance:

$$CI = \hat{Y}_c \pm 1.96 se(\hat{Y}_c)$$

A chosen range of typical survey estimates with their associated standard errors and confidence intervals are presented in the table.

If for example we want to obtain the standard error for the estimate of the number of **unemployed men in the Western Cape for 1997**, find the estimate nearest in value to $\hat{Y}_c = 81091$ from the table (Table 16). Following the conservative approach, the standard error of 100 000 is used, which is 8569. This value is larger than the direct calculated standard error estimate, 7394, but is still acceptable.

Table 16: Table with the standard errors and confidence intervals for the worker data set of the OHS of 1997, according to the official strict definition of unemployment.

Size of Estimate	Standard Error	Lower Confidence Interval	Upper Confidence Interval
1500	645	236	2764
3000	1195	658	5342
5000	1592	1879	8121
10000	2350	5393	14607
30000	4357	21460	38540
50000	5805	38621	61379
70000	7013	56254	83746
100000	8569	83204	116796
300000	15885	268864	331136
500000	21166	458515	541485
700000	25570	649883	750117
1000000	31243	938764	1061236

1300000	36205	1229038	1370962
1500000	39236	1423098	1576902
1700000	42094	1617496	1782504
2000000	46118	1909608	2090392
2300000	49885	2202225	2397775

Advantages:

- This method of presentation makes it very easy for the user to find the standard error, because no calculation is needed.
- The confidence intervals are immediately available to the user.

Disadvantages:

- Often, when the estimate of interest does not match closely with an estimate from the table, it is necessary for the user to use interpolation to find an acceptable estimate of the standard error.
- This presentation method requires that for each new study variable in the survey, a new table must be set up. This can be very time consuming and also take up much space in the publication. It is therefor recommended that this method be used where the survey consists of a limited number of study variables. A good example is the VOC survey where there are only two main study variables, viz. household crimes that are committed against people living together and individual crimes that affect only a single person.
- The table provides estimates of only the same order in size. This may lead to a loss in accuracy in the prediction of the standard error.

4.3 A table with coefficients of relative variation and factor-lines

This presentation method is used by the ABS and is discussed in their Technical Note on Sampling Variability, Appendix D, Household Expenditure Survey (HES) Summary of Results, 1993-1994.

The table consists of the coefficients of relative variation of each study variable at the highest domain level in the survey, e.g. RSA-level, and the necessary factor-lines to be used at lower domain levels, e.g. province, race or gender level (refer to Table 2). The factor-lines are graphically displayed and are used to obtain the necessary adjustment factor with which the given relative standard error should be multiplied to adjust for the smaller sample size of the subclass on which the estimate is based. The coefficients of relative variation are estimated using the fitted model or are calculated directly.

The adjustment factors are calculated by dividing the estimate at a lower domain of interest level by the same estimate at RSA-level and then raised to a power found in the standard error model, e.g.

$$f_a = \left(\frac{\hat{Y}_{Prov}}{\hat{Y}_{RSA}} \right)^{Power} \tag{23}$$

where f_a denotes the adjustment factor, $\hat{Y}_{RSA}$ the estimate at RSA-level and $\hat{Y}_{Prov}$ the estimate at a lower level, e.g. province-level.

This procedure can be justified mathematically as follows: To estimate the natural logarithm of the coefficient of relative variation at RSA-level, the formula is:

$$\ln(cv(\hat{Y}_{RSA})) = a + b\ln(\hat{Y}_{RSA})$$

and at a lower level, e.g. at province-level:

$$\ln(cv(\hat{Y}_{Prov})) = a + b\ln(\frac{\hat{Y}_{Prov}}{\hat{Y}_{RSA}} \times \hat{Y}_{RSA})$$

$$\therefore cv(\hat{Y}_{Prov}) = e^{a+b\ln(\hat{Y}_{RSA})+b\ln(\frac{\hat{Y}_{Prov}}{\hat{Y}_{RSA}})}$$

$$= e^a \times e^{b\ln(\hat{Y}_{RSA})} \times e^{b\ln(\frac{\hat{Y}_{Prov}}{\hat{Y}_{RSA}})}$$

$$= e^a \times (\hat{Y}_{RSA})^b \times \left(\frac{\hat{Y}_{Prov}}{\hat{Y}_{RSA}} \right)^b$$

$$= cv(\hat{Y}_{RSA}) \times f_a \tag{24}$$

The following steps must be followed to find the estimated coefficient of relative variation of interest:

- Obtain the estimated value of the study variable from the published table.
- Obtain the estimated coefficient of relative variation for this study variable at RSA-level and its factor-line from the table (Table 17 in the case of the OHS of 1997).
- Read off the adjustment factor for the estimate of interest and the specific factor-line from the factor-line graph which is provided (Figure 10 in the case of the OHS of 1997).
- The estimated coefficient of relative variation for the estimate at a lower level is calculated as: $cv_{lowerlevel} = f_a \times cv_{RSA}$

4.3.1 Example:

To compare the standard error given by this presentation method with the standard errors of the previous methods, again the example of the estimated number of **unemployed males in the Western Cape** from the worker data set of the OHS of 1997 is used. The estimate of interest is: $\hat{Y}_{WC \times M} = 81091$. From Table 17 we obtain the

coefficient of relative variation for the study variable at RSA-level: the number of unemployed people in the RSA:

$$cv(\hat{Y}_{RSA}) = 0.0212$$

The factor-line to use is: **I** and from Figure 10 on page 49 we find the factor for $\hat{Y}_{WC\times M} = 81091$, is: $f_a = 4.4$

$$\therefore cv(\hat{Y}_{WC\times M}) = 4.4 \times 0.0212$$

$$cv(\hat{Y}_{WC\times M}) = 0.0933$$

To calculate the estimated standard error, the coefficient of relative variation must be multiplied with the estimate, $\hat{Y}_{WC\times M}$:

$$se(\hat{Y}_{WC\times M}) = 0.0933 \times 81091$$

$$se(\hat{Y}_{WC\times M}) = 7565$$

This value compares well with the direct calculated standard error estimate, 7394, for the same estimate. In the same way the standard error estimate for $\hat{R}$ can be obtained.

Table 17: Table with coefficients of relative variation at RSA level for the worker data set and the household data set of the October Household Survey of 1997, and factor-lines to derive the relative standard errors at lower levels of the domains of interest.

Study Variable from survey	Coefficient of Relative Variation of $\hat{Y}_c$ = estimated number	Coefficient of Relative Variation of $\hat{R}$ = estimated ratio	Factor-lines At lower levels, e.g. province-, race-, gender-level, etc.
OHS 1997 – Worker data set			
Unemployed in RSA	0.0212	0.0178	I
OHS 1997 – Household data set			
Dwelling Types			
Households with a formal house or brick structure on a separate stand or yard in RSA	0.0242	0.0188	A
Households with traditional dwelling, hut, structure, made of traditional materials	0.0444	0.0363	B
Households living in flats, apartment in block of flats	0.0676	0.0573	C
Town-, cluster-, semi-detached house (simplex, duplex, or triplex)	0.0808	0.0695	D
Households with an informal dwelling, shack, in the back yard	0.2137	0.2	E

Households with an informal dwelling, shack, NOT in the back yard, e.g. in an informal squatter settlement	0.0988	0.0865	F
Room in hostel, compound for workers provided by employer or municipality	0.0923	0.0803	G
Main source of Water	$cv(\hat{Y}_c)$	$cv(\hat{R})$	Factor-line
Piped (tap) water, in dwelling	0.0257	0.0218	A
Piped (tap) water, on site or in yard	0.0327	0.0281	B
Public tap	0.0357	0.0308	B
Water-Carrier, tanker	0.111	0.101	E
Borehole on site	0.1079	0.098	H
Borehole: off site, communal	0.0697	0.0623	F
Rain-water tank on site	0.1845	0.1717	E
Flowing water, stream	0.0516	0.0453	B
Dam, pool, stagnant water	0.0907	0.0817	H
Well	0.1126	0.1024	D
Spring	0.0846	0.076	H

Advantage:

- This presentation method is fairly easy for the user to apply.

Disadvantages:

- Many study variables from the survey require many table entries.
- To set up the table requires much work and time.
- The method requires the user to be familiar with the use of graphs and to do some simple calculations to obtain the estimated value of the standard error.

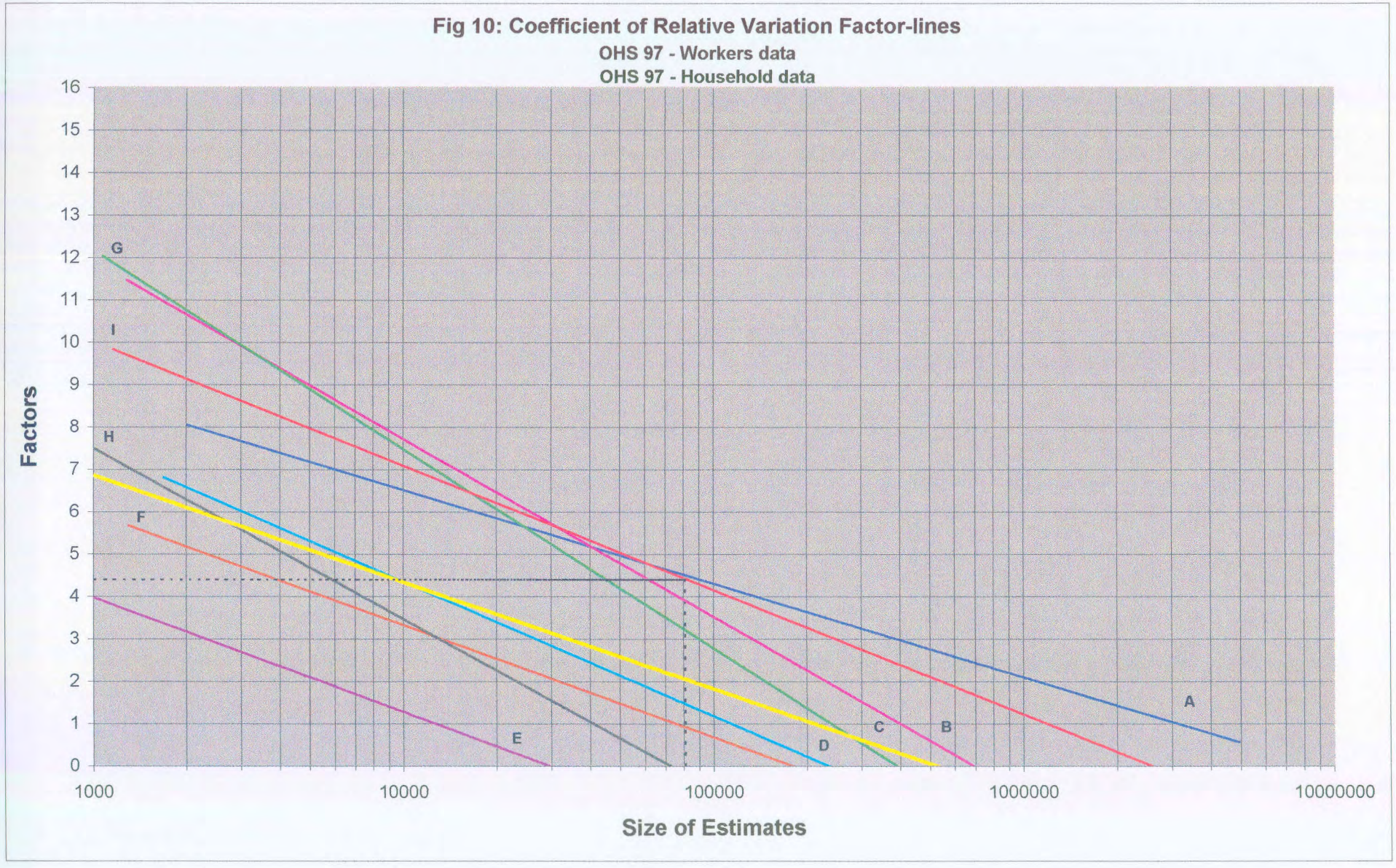


Fig 10: Coefficient of Relative Variation Factor-lines
OHS 97 - Workers data
OHS 97 - Household data
Factors
Size of Estimates
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
1000
10000
100000
1000000
10000000
A
B
C
D
E
F
G
H
I

4.4 Formulas and Graphs

The model for the coefficient of relative variation for each study variable from the survey is published and can also be graphically presented. The user only needs to insert the value of the estimate of interest into the model or has the option to read off the coefficient of relative variation from the published graph. The necessary conversion formula to calculate the standard error from the coefficient of relative variation must also be given with an example that explains to the user how the formulas and the graphs should be used.

The formulas to calculate confidence intervals can also be included and explained to the user as indicated below. This comment is applicable to all the previous presentation methods as well.

4.4.1 Example

Returning to the example that has already been used, the estimated number of **unemployed men in the Western Cape** from the 1997 OHS worker data set is 81091.

The model to use is:

$$\begin{aligned}\ln\left(cv(\hat{Y}_c)\right) &= 2.588 - 0.4382\ln(\hat{Y}_c)\\ &= 2.588 - 0.4382\ln(81091)\\ &= -2.3651\end{aligned}$$

To convert this value into the standard error, the following formula is used:

$$\begin{aligned}se(\hat{Y}_c) &= \exp\left(\ln(cv(\hat{Y}_c))\right)\times\hat{Y}_c\\ &= 0.09394\times 81091\\ &= 7618\end{aligned}$$

Alternatively, the formula $se(\hat{Y}_c) = 13.303\times\hat{Y}_c^{\,0.5618}$ as derived on page 28 can be used which will give the same standard error (7618) when $\hat{Y}_c = 81091$ is substituted into this model.

If the graph is used, we find the coefficient of relative variation for $\hat{Y}_c = 81091$ is: $cv(\hat{Y}_c) = 0.095$ (see the dotted line on Figure 11, page 51)

To calculate the standard error: $\begin{aligned}se(\hat{Y}_c) &= cv(\hat{Y}_c)\times\hat{Y}_c\\ &= 0.095\times 81091\\ &= 7704\end{aligned}$

Although not exactly the same, this value compares well to the standard error calculated with the formula above.

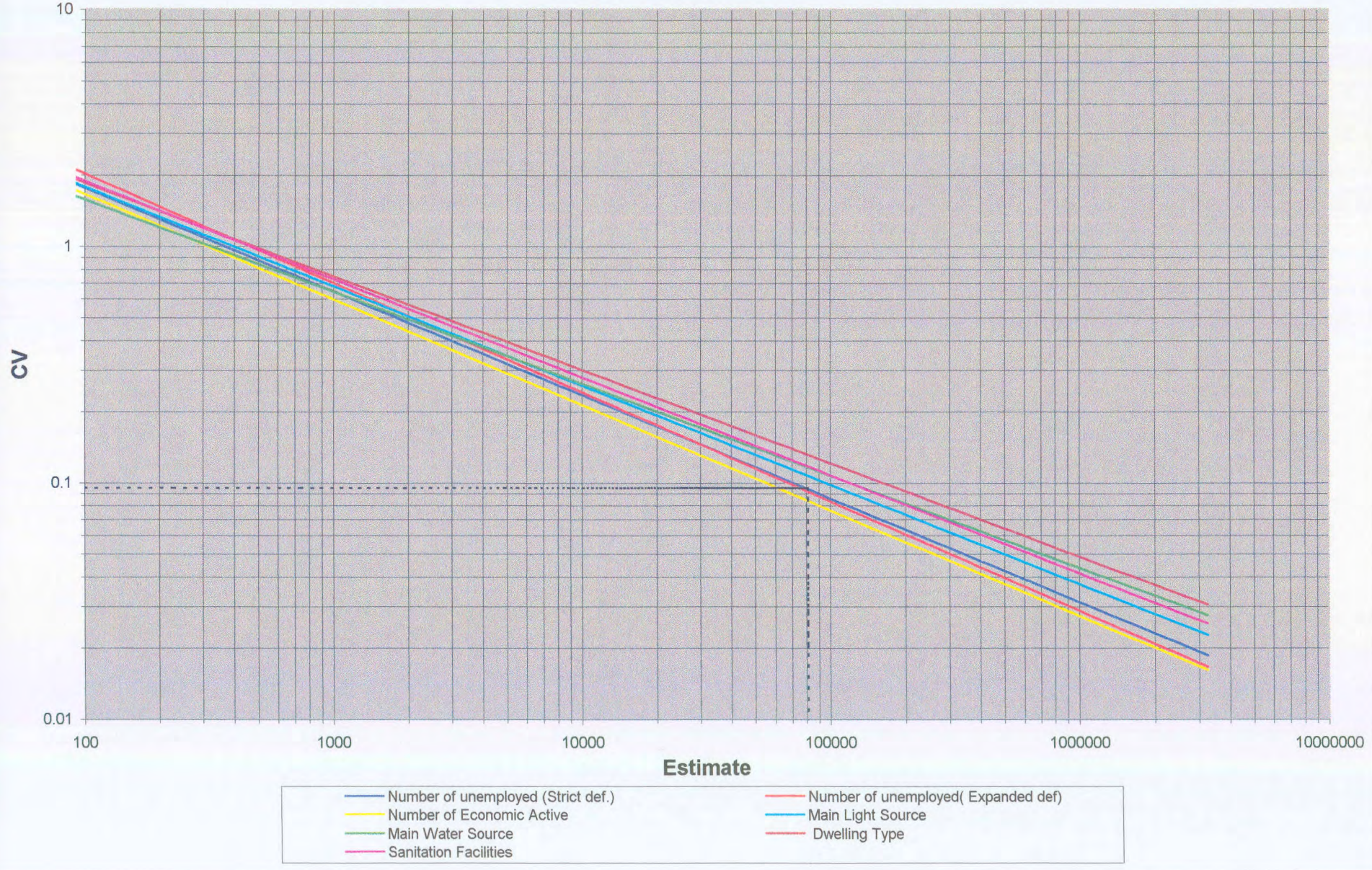
To calculate the 95% confidence interval for this estimate:

$$\begin{aligned}CI &= \hat{Y}_c \pm 1.96 se(\hat{Y}_c)\\ &= 81091 \pm 1.96\times 7704\\ &= [65991\,;\,96191]\end{aligned}$$

Fig 11: The estimated coefficient of relative variation

OHS 97 - Workers data

OHS 97 - Household data



If the standard error and confidence interval for $\hat{R}$ are required, $\hat{Y}_c$ must be replaced with $\hat{R}$ in the above formulas where applicable.

Advantages:

- The formulas presented are very easy to use for the statistician and will result in getting a better estimated value for the standard error.
- The method in graphical form is very easy to understand and to be used by the general user.
- This method is also very space efficient.

Disadvantage:

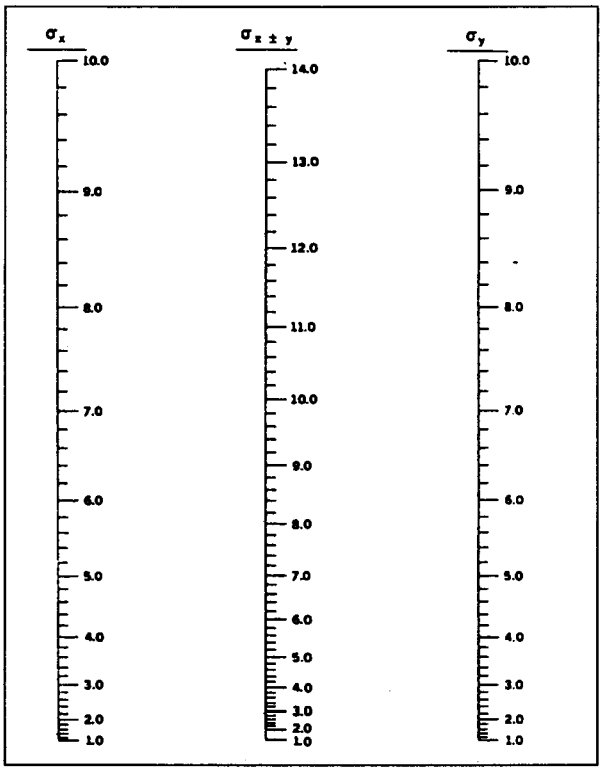
- The method requires users to be familiar with the use of graphs and / or formulas.

4.5 Nomogram

A nomogram is a graphical presentation for mathematical functions consisting of more than one independent variable. The model, $\ln(cv(\hat{Y}_c)) = a + b\ln(\hat{Y}_c)$, is a simple straight line with only one independent variable. It will serve no purpose to construct a nomogram for this model.

However, a nomogram can be an extremely valuable tool to facilitate calculations. For example, it may be necessary to test whether estimates of the unemployment rate obtained in independent cross-sectional surveys such as the OHS of 1995 and the OHS of 1996, differ significantly. This will require an estimate of the standard error of the difference between the estimated values.

Fig 12: Nomogram: Standard error of sum or difference



Instructions to use the nomogram: Let $\hat{X}$ and $\hat{Y}$ be two independent estimates of X and Y respectively. $\hat{X}+\hat{Y}$ is an estimate of the sum and $\hat{X}-\hat{Y}$ is an estimate of the difference of $\hat{X}$ and $\hat{Y}$. The nomogram can be used to approximate the standard errors of $\hat{X}+\hat{Y}$ and $\hat{X}-\hat{Y}$ by following the steps:

- Find the point on the σ_x- scale that corresponds to the estimated standard error of $\hat{X}$ and the point on the σ_y- scale that corresponds to the estimated standard error of $\hat{Y}$
- The scales may be read in any unit (tenths, thousands, millions) as long as the same unit is used on all the scales
- Connect these points on the σ_x- scale and the σ_y- scale by a straight line. The value where the line crosses the $\sigma_{x\pm y}$- scale is the estimated standard error of $\hat{X}+\hat{Y}$ and $\hat{X}-\hat{Y}$.

If for example $se(\hat{X})=6.75$ and $se(\hat{Y})=4.7$, a straight line connecting these points, crosses the $\sigma_{x\pm y}$- scale at about 8.25 while an exact computation gives 8,225 (Gonzalez, Ogus, Shapiro and Tepping; 1975).

To test whether an observed difference between the unemployment rates, $\hat{R}_1$ and $\hat{R}_2$, obtained in different OHSs is statistically significant, the 95% confidence interval for the difference must be calculated, viz.:

$$\left((\hat{R}_1-\hat{R}_2)-1.96se(\hat{R}_1-\hat{R}_2)\ ;\ (\hat{R}_1-\hat{R}_2)+1.96se(\hat{R}_1-\hat{R}_2)\right) \tag{25}$$

If this interval does not include the value 0, then the estimated unemployment rates $\hat{R}_1$ and $\hat{R}_2$, differ significantly at the 5% level of significance (using two-sided testing).

Nomograms require more effort to set up, but are very easy to use by the survey user.

Considering the advantages and disadvantages of the different presentation methods (as given in the discussion of each method), the following conclusions are reached.

The extent of the survey plays a very important role in the choice of a presentation method of standard error estimates in a survey report. For a survey which involves many study variables and many domains of interest (as in the case of the OHSs) graphs and formulas would be most cost effective to present the standard error estimates. On the other hand, tables would take up too much space in the publication and much time to set up. However, when the survey involves only a few study variables, tables are practical and effective.

The background of the survey user will also influence the choice of a presentation method. If they are unfamiliar with the use of graphs and formulas, a table with standard error estimates according to different sizes of the estimates, is probably the easiest method to understand and use. Thus, the circumstances of a specific survey study will determine the choice of the most suitable presentation method.

5. Concluding remarks

This research project addressed a very common problem experienced whenever a survey of mention-able size, which involves many different study variables and many different domain subclasses of interest, is being conducted: how to estimate and present the standard errors in the survey report without taking up too much valuable time and space in the publication.

The first part of the research project examined the feasibility of modeling the standard errors of estimated population parameters or characteristics of the study variable of interest, using one mathematical model over all the different domains of interest, whether the domain is a cross-class, segregated class or a mixed-class. The results were satisfying in showing that mathematical modeling of standard errors can be done with great success and the same model seems to fit equally well over all the different domain subclasses.

The research methodology consists of the testing of different standard error models used by countries like the USA and Australia for the purpose of modeling standard errors in survey reports. These models were fitted to typical South African data sets (OHSs of 1995, 1996 and 1997 and the VOC of 1998) by means of Least Squares regression modeling via SAS INSIGHT. The obtained R^2-values of the different models fitted, were compared with each other in order to find the best suitable model for the data sets considered. The model discovered to give the best fitting results was derived and is still in use by the Australian Bureau of Statistics. This model differs only in the respective model parameters for the estimated total and the estimated ratio and can therefor easily be displayed on a single graph, which makes presentation of the models in the survey report easier. Although certain cases were identified that are likely to give rise to outliers and consequently may influence the model fitting results negatively, there are solutions proposed for each of these cases which can be followed in order to obtain the best possible fitting results.

The second part of the research project focused on the finding of a practical and efficient presentation method of standard errors in the published survey report. Several methods that comply to these requirements were identified and introduced, but the choice of the best method depends to a large extent on the extent of the survey (i.e. the number of different study variables and the number of different domains involved) and the background of the user of the report (i.e. the familiarity of the survey user with the use of formulas and graphs, etc.).

The presentation method that is identified to require the least trouble to include in the survey report and that is also very easy to apply, is to include the fitted model in terms of a formula in the survey report. The survey user only has to substitute the estimated value of the study variable of which the standard error is required, into the given formula or has the alternative to read off the estimated coefficient of relative variation from the graph that is also included in the survey report. The formula to calculate the confidence intervals must also be included in the survey report and provide the survey user with a user-friendly and cost-effective method to examine the precision of the survey results.

The conclusion reached is that the research results positively support the use of mathematical modeling to estimate standard errors and can also be used very efficiently in the presentation of standard errors in the survey report.

6. Glossary

Coefficient of relative variation (cv)

It is the estimated standard error given as a proportion of the estimated value and can be defined as: $cv = \frac{se(\hat{Y})}{\hat{Y}}$

where $se(\hat{Y})$ denotes the estimated standard error and $\hat{Y}$ the estimated total.

Complex Sampling (CS)

Can be described as *multistage stratified cluster sampling* consisting of different sampling stages in which any of the four probability sampling methods, viz. Simple Random Sampling (SRS), Systematic Sampling (SS), Cluster Sampling and Stratified Sampling (STR) can be used.

multi stage: more than one sampling stage exist
stratified: the population is first divided into non-overlapping subpopulations called strata; sampling is done independently within each stratum
cluster: naturally formed subgroups where each population element belongs to one and only one cluster
Note the difference between strata and clusters: strata contain clusters completely and clusters are drawn from strata.

Design effect (deff)

Ratio of the actual sampling variance, taking into account the complexity of the sampling design, to the variance of the same sample size under assumptions of SRS (Kish; 1965).
Thus, *deff* measures the combined effect of stratification and clustering on precision, as measured by the variance, compared to the variance obtained by the direct application of SRS.

$$deff = \frac{Variance\ of\ an\ estimate\ under\ CS}{Variance\ of\ an\ estimate\ under\ SRS}$$

Domain

Population subgroups or subclasses that are formed by classifying according to one or more categorical predictors such as gender, age groups, race, etc.

Domain Subclass types:

Cross-classes

A type of subclass that cuts smoothly across the clusters and strata. These classes are more or less uniformly distributed across the whole population. It includes subclasses by age and gender.

Segregated classes

A term used to indicate subclasses that are completely segregated into separate classes. The whole cluster either belongs or does not belong to the subclass. Examples are provinces, urban / rural areas, geographical areas, etc.

Mixed classes

These classes are less well distributed than cross-classes, but are not completely separated into segregated classes. Examples are ethnic or racial subclasses, occupational and other socio-economic classes.

Economically active population

Include the *workers* (see page 57) and the unemployed with age between 15 and 65 years.

Enumerated Area (EA)

A term used by Statistics South Africa for a well-demarcated geographic area containing in general between 100 and 250 households, depending on the type of area, and enumerated by an enumerator in the area.

Estimator

An *estimator* of a population parameter or characteristic is a mathematical formula used to estimate population parameters or characteristics. The numerical value obtained by using the *estimator* for the actual sample is called the *estimate*.

Population parameters or characteristics

Are functions of the population values Y_k of the study variable y and may also include population values X_k of an auxiliary variable x. Typical population parameters or characteristics are the total, the ratio and the median. (Lethonen and Pahkinen; 1995)

Precision of an estimator

The 95% confidence interval for the estimated population parameter $\hat{Y}$ with estimated standard error $se(\hat{Y})$ is being calculated by: $[\hat{Y} - 1.96se(\hat{Y})\ ;\ \hat{Y} + 1.96se(\hat{Y})]$, assuming that $\hat{Y}$ is distributed according to the standard normal distribution. The quantity $1.96se(\hat{Y})$ is called the precision of the estimate.

Sampling design

A sample is a subset of the finite population U. The *sampling design* refers to the specific probability-sampling scheme used to draw the sample.

Simple Random Sampling (SRS)

Each element in the population receives the same probability to be selected in the sample at each draw of an element in the population.

Sampling without replacement

After an element has been drawn from the population, it is not present in the population anymore. Consequently such elements can appear only once in the sample.

Sampling with replacement

Elements are replaced in the population after each draw and consequently can be drawn more than once in the sample.

Sampling units defined according to sampling stage:

Primary sampling unit (PSU)

The sampling units that are drawn in the first sampling stage are called primary sampling units.

Ultimate sampling unit (USU)

The sampling units that are drawn in the last sampling stage are called ultimate sampling units.

Study variable

The variable of interest for which measurements are recorded in a sample survey where the sample is drawn according to a specific probability-sampling scheme. If a finite population of N elements is denoted by $U = \{1,...,k,...,N\}$, then the *study variable* is denoted by y, with unknown population values $Y_1,...,Y_k,...,Y_N$. Usually there is also an *auxiliary study variable* x, with unknown population values $X_1,...,X_k,...,X_N$ (Lethonen and Pahkinen; 1995)

Unemployment definitions used by Statistics South Africa:

Official or strict definition of unemployment

Statistics South Africa uses the following definition of unemployment as the official definition. The unemployed are those people within the economically active population who: (a) did not work during the seven days prior to the interview, (b) want to work and are available to start work within a week of the interview and (c) have taken active steps to look for work or to start some form of self-employment in the four weeks prior to the interview

Expanded definition of unemployment

The same definition as above but, without the requirement of criterion (c).

Variance

The estimates of a population parameter vary from sample to sample. This variation that exists because of the probabilistic nature of a sample, is called the sampling variance. The sampling error is measured by the standard error = $\sqrt{Variance}$.

Workers

People who have worked or had a job in the 7 days prior to the interview.

7. References

ABS see The Australian Bureau of Statistics

Bieler, G. S., and Williams, R. L., 1990. **'Generalized Standard Error Models for Proportions in Complex Design Surveys'** in *Proceedings of Section on Survey Research Methods of the American Statistical Association*, 272-277

Cochran, 1977. '**Sampling Techniques**', Third Edition. John Wiley & Sons, New York

Cox, B. G., Jang, D., Edson, D., 1993. **'Sampling Errors for SESTAT and its Component Surveys: 1993',** Mathematica Policy Research, Inc.

Finamore, J. M., 1999. **'Generalized Variance Parameters for the 1997 National Survey of College Graduates'**, U. S. Bureau of the Census, Demographic Statistical Methods Division, Health Surveys and Supplements Branch

Ghangurde, P. D., 1981. **'Models for Estimation of Sampling Errors'** *Survey Methodology*, **7**, 177-191

Gonzalez, M. E., Ogus, J. L., Shapiro, O. G., and Tepping, B. J., 1975. **'Standards for Discussion and Presentation of Errors in Survey and Census Data'** in *Journal of the American Statistical Association*, **70**(351), Part II, 5-23

Johnson, E. G., and King, B. F., 1987. **'Generalized Variance Functions for a Complex Sample Survey'** in *Journal of Official Statistics*, **3**, 235-250

Kalton, G., 1977. **'Practical Methods for Estimating Survey Sampling Errors'** in *Bulletin of the International Statistical institute*, **47**, 495-514

Kish, L., 1965. **'Survey Sampling'**, John Wiley & Sons, New York

Lepkowski, 1998. '**Presentation of Sampling Errors'**, Methods *of Survey Sampling / Applied Sampling* – Lecture at Statistics South Africa

Lethonen, R., and Pahkinen, E. J., 1995. **'Practical Methods for Design and Analysis of Complex Sample Surveys'**, Revised Edition. John Wiley & Sons

Neethling, A., Stoker, D. J., and Eiselen, R., 1997. **'Complex Sampling and the Analysis of Complex Sample Data'**, Workshop on Official Statistics, Presented at the 1997 Annual Conference of the South African Statistical Association, 1-37

Stoker, 1999. **'Notes on sampling theory',** *Design and Analysis of Complex Samples*, A.1,1-11

The Australian Bureau of Statistics, 1993-1994. **'Technical Note on Sampling Variability'** in *ABS – HES Summary of Results*, Appendix D, 43-49

The Australian Bureau of Statistics, 1997. **'Household Collection Support Standard Error Manual'**, Section 3, 29-41

Valliant, R., 1987. **'Generalized Variance Functions in Stratified Two-Stage Sampling'** in *Journal of the American Statistical Association*, **82**, 499-508

Verma, V., 1982. '**The Estimation and Presentation of Sampling Errors**' in *World Fertility Survey Technical Bulletins*, **11**

8. Appendix - A

Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated number of unemployed according to the strict definition of unemployment. (Source: OHS 97 – Workers file)

Model: $\ln(cv(\hat{R})) = 2.7087 - 0.45851\ln(\hat{Y}_c)$

Fig 1:

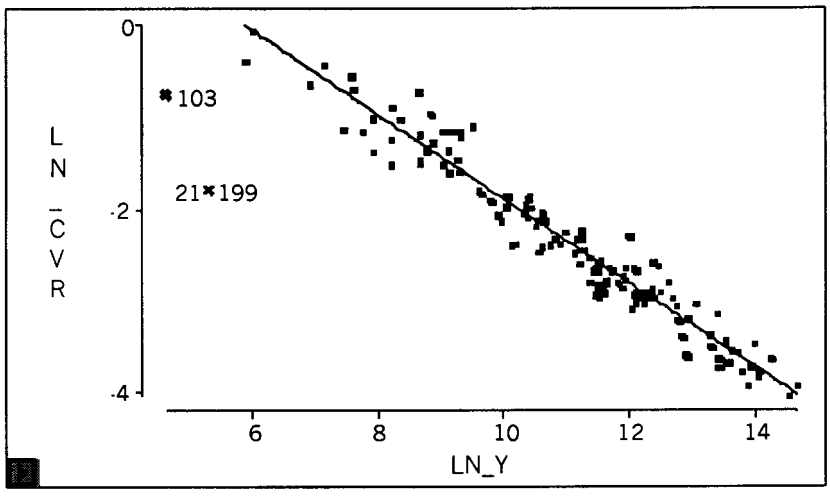


Table 1:

Parametric Regression Fit								
			Model		Error			
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
	1	1	142.1090	194	0.0411	0.9469	3461.1108	0.0001

Table 2:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob > |T|	Tolerance	Var Inflation
INTERCEPT	1	2.7087	0.0887	30.5223	0.0001	.	0
LN Y	1	-0.4585	0.0078	-58.8312	0.0001	1.0000	1.0000

Fig 2:

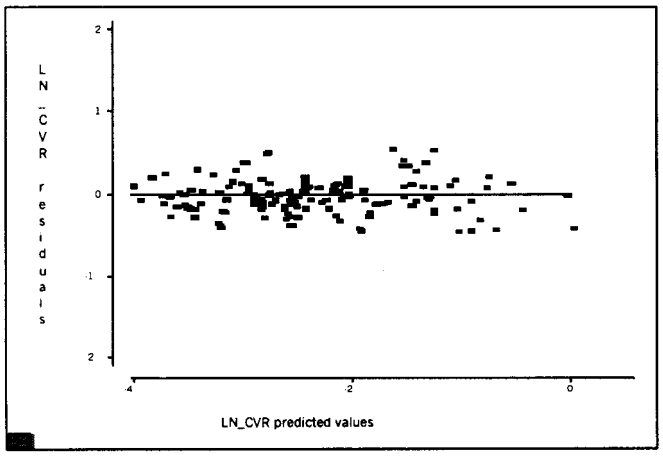
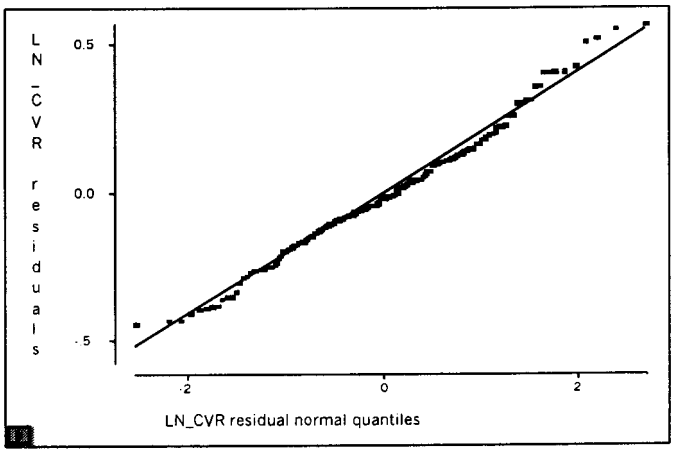


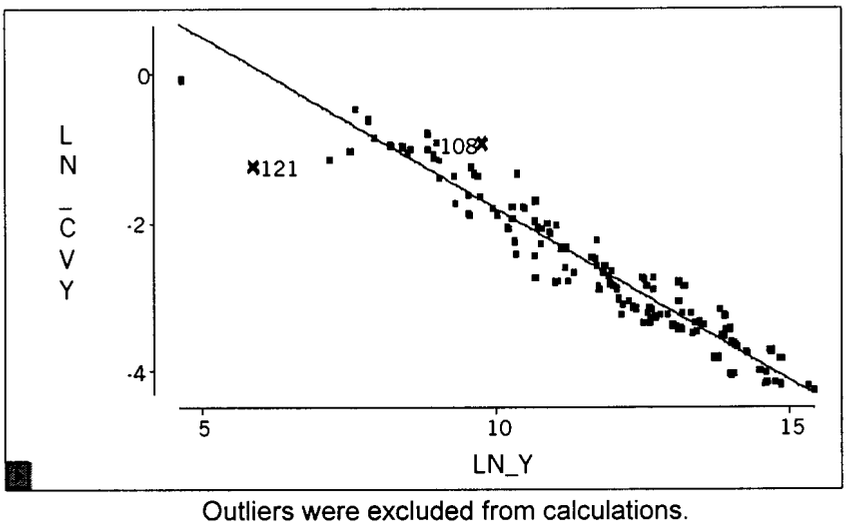
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population total unemployed in South Africa, according to the expanded definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$. (Source: OHS 97 – Household file)

Model: $\ln(cv(\hat{Y}_c)) = 2.8358 - 0.4623\ln(\hat{Y}_c)$

Fig 1:



Outliers were excluded from calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
_____	1	1	136.2614	160	0.0591	0.9351	2305.5189	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.8358	0.1157	24.5031	0.0001	.	0
LN Y	1	-0.4623	0.0096	-48.0158	0.0001	1.0000	1.0000

Fig 2:

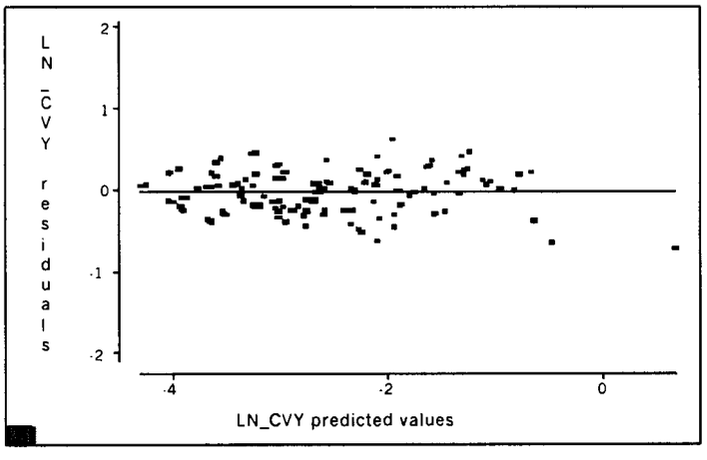
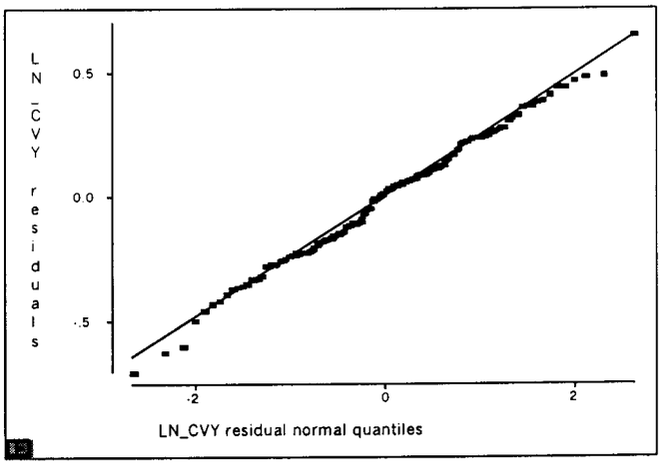


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio unemployed in South Africa, according to the expanded definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated number of unemployed according to the expanded definition of unemployment. (Source: OHS 97 – Workers)

$$\textbf{Model: } \ln(cv(\hat{R})) = 2.9865 - 0.4887 \ln(\hat{Y}_c)$$

Fig 1:

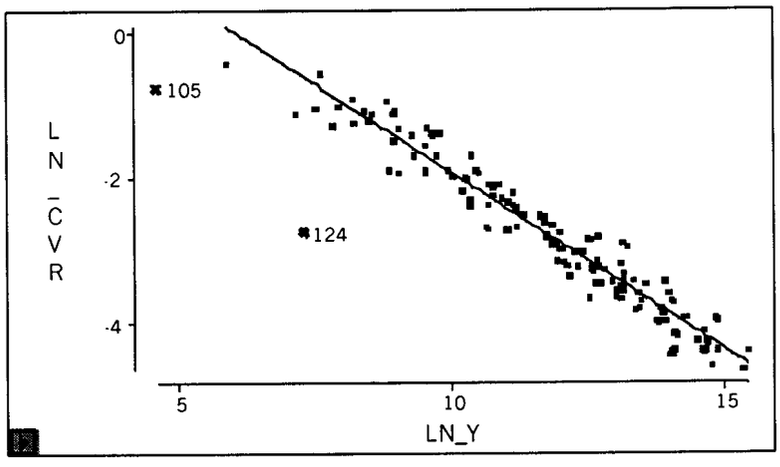


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	149.3041	161	0.0610	0.9383	2449.1061	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.9865	0.1186	25.1795	0.0001	.	0
LN_Y	1	-0.4887	0.0099	-49.4884	0.0001	1.0000	1.0000

Fig 2:

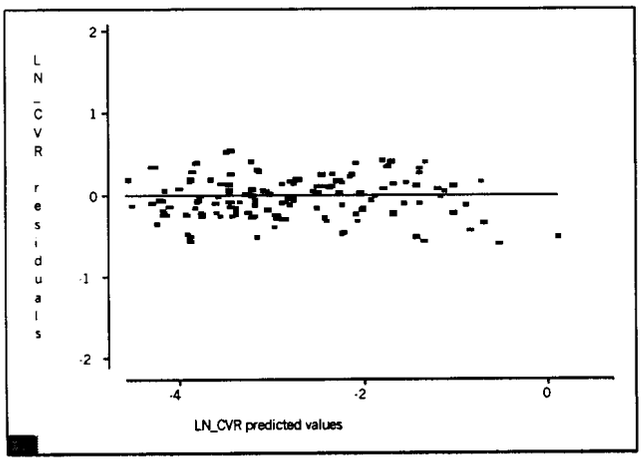
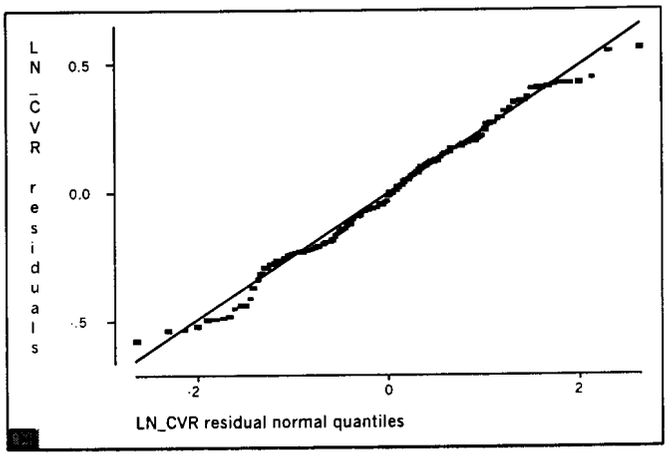


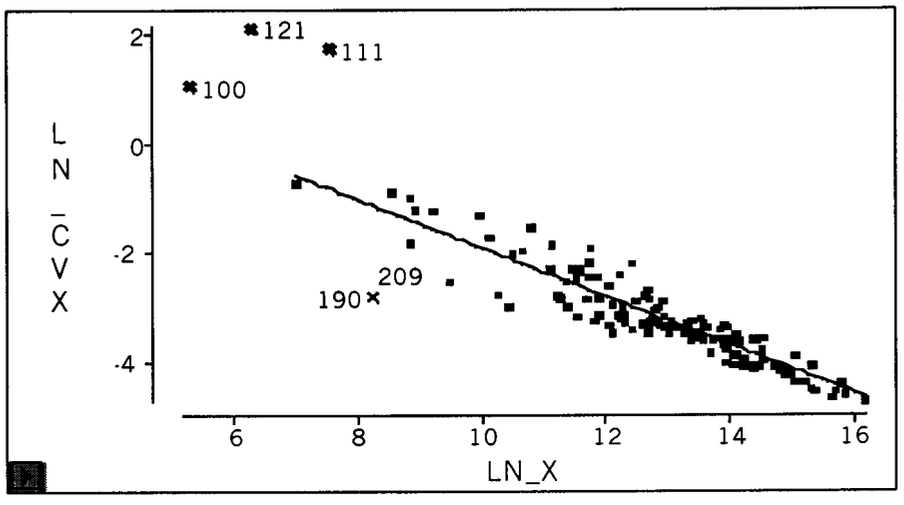
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of economic active people in South Africa as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 97 – Workers)

Model: $\ln(cv(\hat{Y}_c)) = 2.5627 - 0.4462\ln(\hat{Y}_c)$

Fig: 1



All outliers have been excluded from the calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	99.4425	195	0.0833	0.8597	1194.4344	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5627	0.1699	15.0844	0.0001	.	0
LN X	1	-0.4462	0.0129	-34.5606	0.0001	1.0000	1.0000

Fig 2:

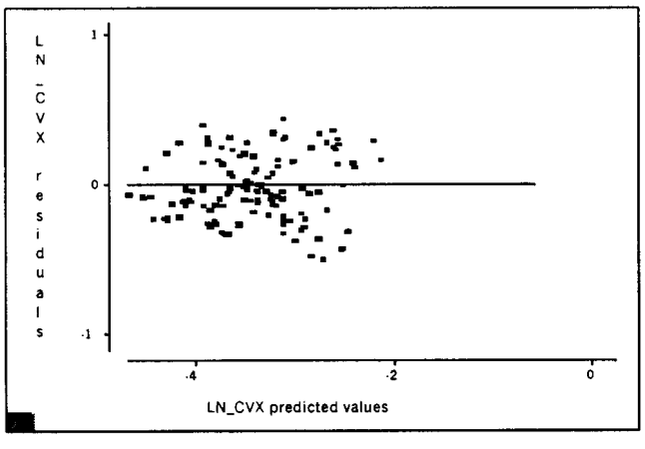
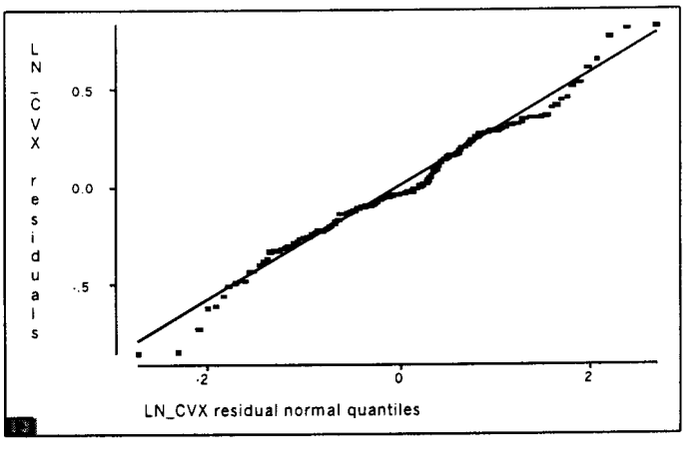


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa according to different lighting sources, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 97 – Household file)

Model: $\ln(cv(\hat{Y}_c)) = 2.5152 - 0.4202\ln(\hat{Y}_c)$

Fig: 1

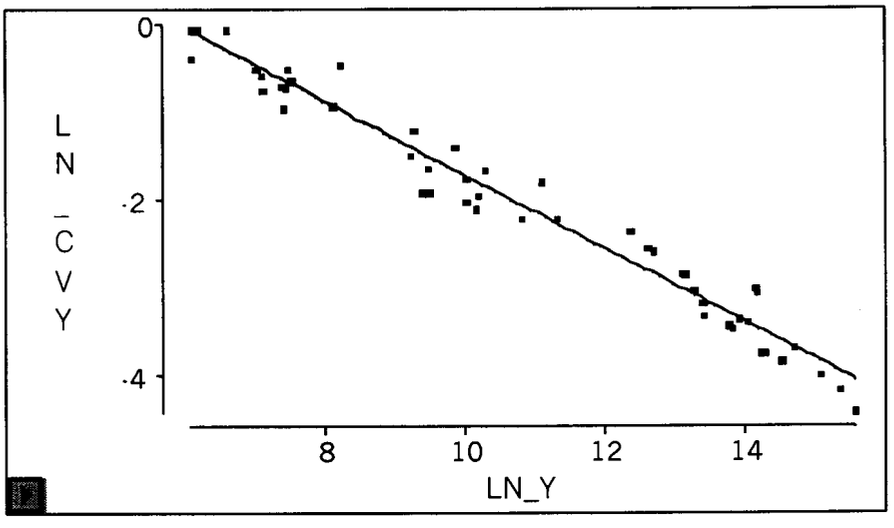


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	83.3974	51	0.0576	0.9660	1449.0311	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5152	0.1250	20.1148	0.0001	.	0
LN_Y	1	-0.4202	0.0110	-38.0661	0.0001	1.0000	1.0000

Fig 2:

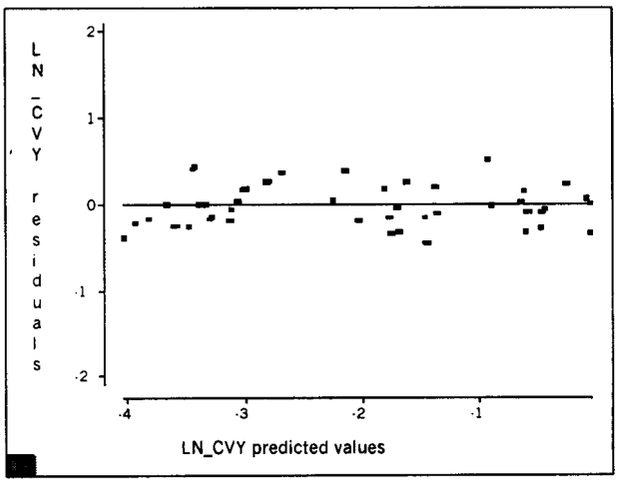
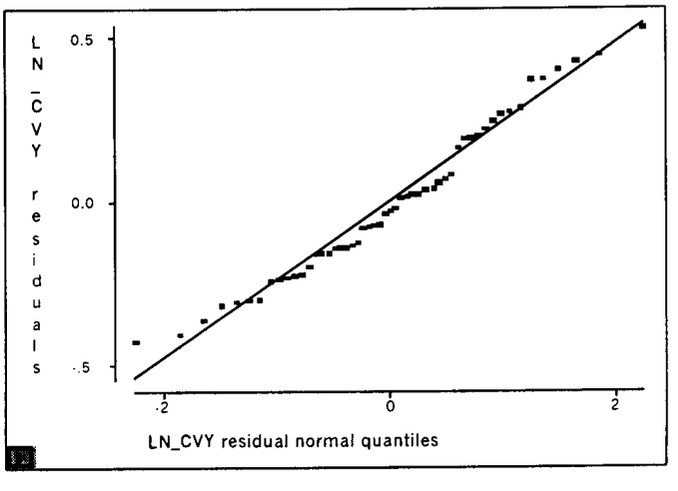


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different lighting sources, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated number of households with different lighting sources. (Source: OHS 97 – Workers file)

Model: $\ln(cv(\hat{R})) = 2.772 - 0.4642 \ln(\hat{Y}_c)$

Fig: 1

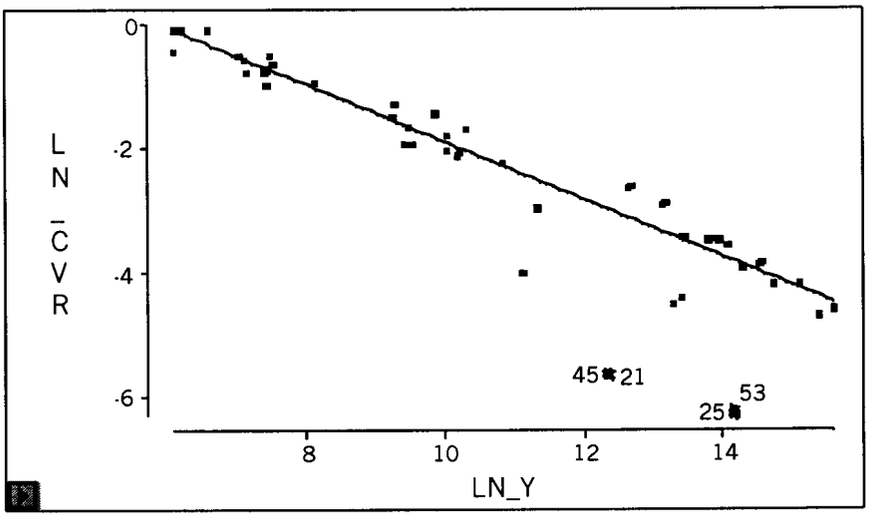


Table 1:

Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
———	1	1	94.5395	46	0.1526	0.9309	619.5069	0.0001

Table 2:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob > |T|	Tolerance	Var Inflation
INTERCEPT	1	2.7720	0.2090	13.2662	0.0001	.	0
LN Y	1	-0.4642	0.0187	-24.8899	0.0001	1.0000	1.0000

Fig 2:

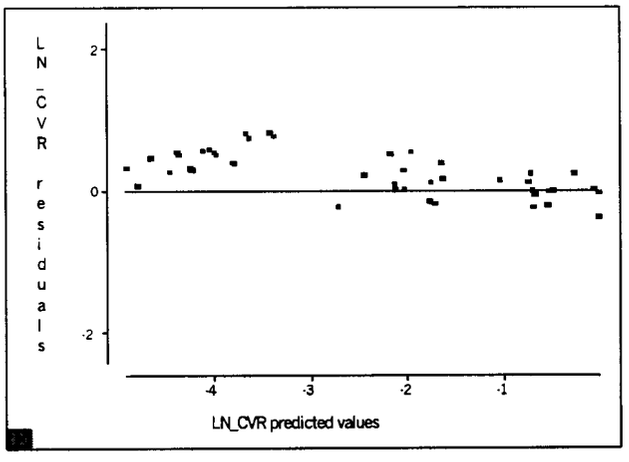
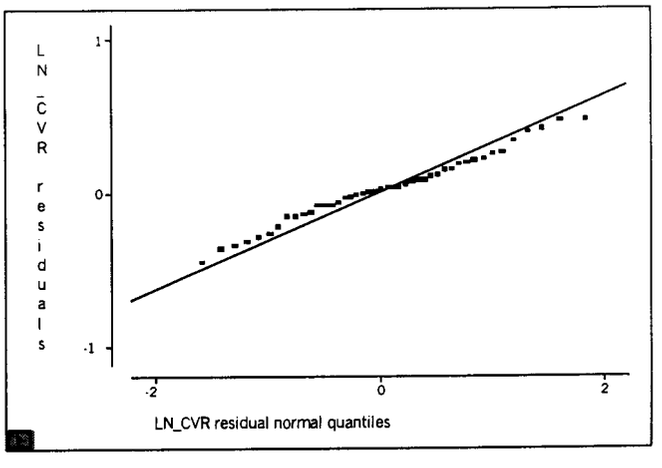


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa according to different water sources, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 97 – Household file)

Model: $\ln(cv(\hat{Y}_c)) = 2.2501 - 0.3896\ln(\hat{Y}_c)$

Fig: 1

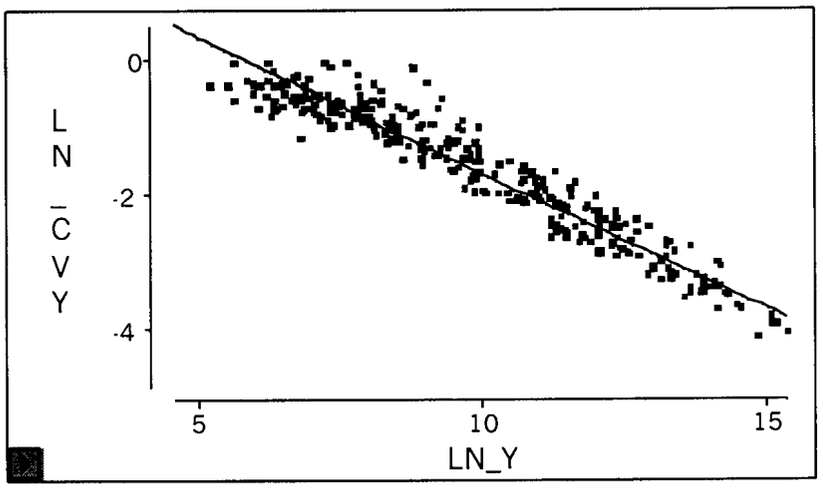


Table 1:

Parametric Regression Fit

		Model		Error			
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob>F
	1	1	475.4534	463	0.0811	0.9268	5865.5286	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.2501	0.0499	45.1316	0.0001	.	0
LN_Y	1	-0.3896	0.0051	-76.5867	0.0001	1.0000	1.0000

Fig 2:

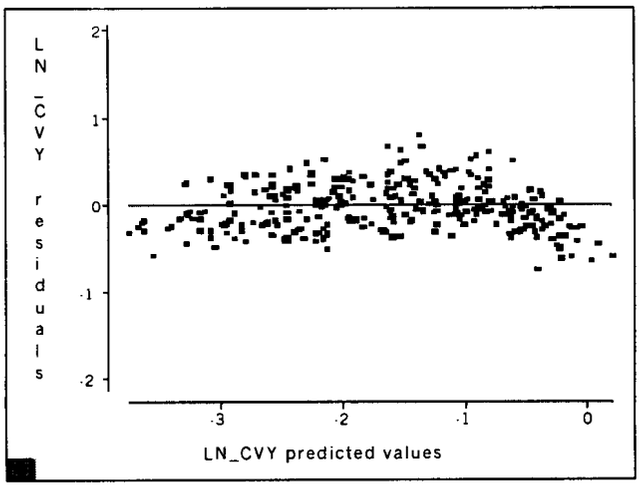
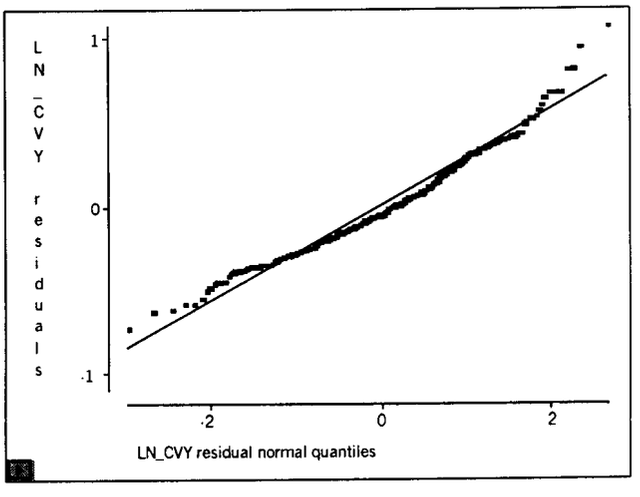


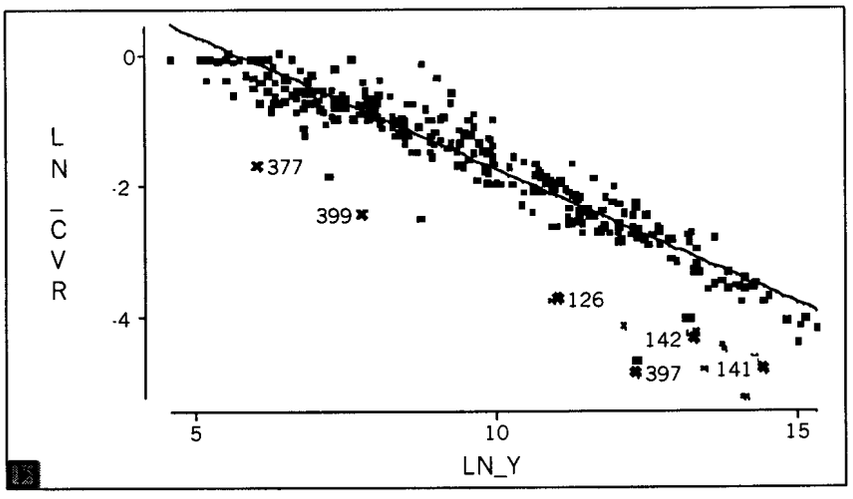
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different water sources, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated population number of households with different water sources.
(Source: OHS 97 - Household file)

$$\textbf{Model: } \ln(cv(\hat{R})) = 2.3443 - 0.4067\ln(\hat{Y}_c)$$

Fig: 1



Observations identified as outliers, indicated with x, have been excluded.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	503.9124	455	0.1134	0.9071	4442.5419	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.3443	0.0596	39.3541	0.0001	.	0
LN Y	1	-0.4067	0.0061	-66.6524	0.0001	1.0000	1.0000

Fig 2:

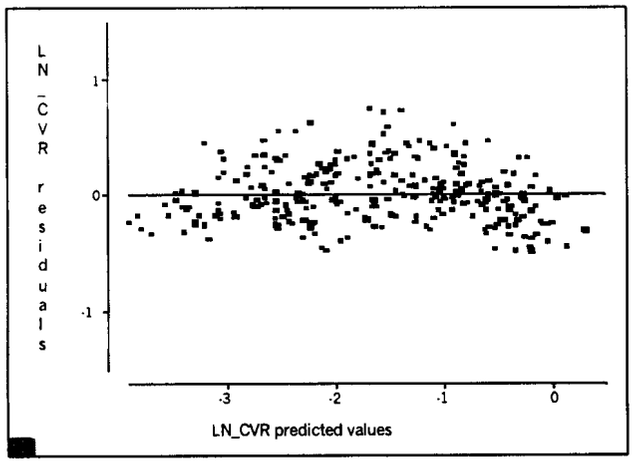
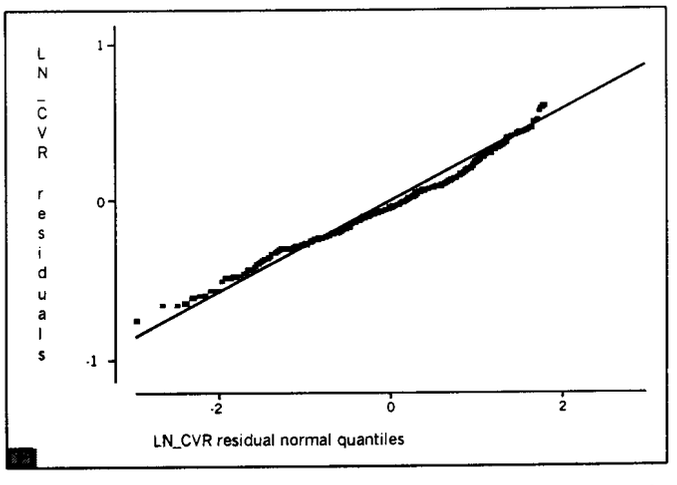


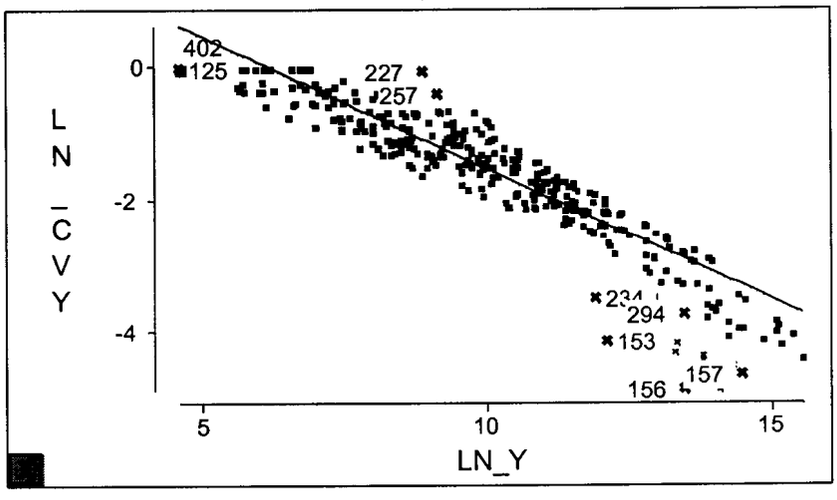
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa according to different dwelling-types, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 97 – Household file)

Model: $\ln(cv(\hat{Y}_c)) = 2.4389 - 0.3955\ln(\hat{Y}_c)$

Fig: 1



All the outliers, indicated with x, were excluded from the calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	DF	Model Mean Square	DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	357.6910	437	0.0989	0.8922	3615.5251	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.4389	0.0661	36.9163	0.0001	.	0
LN_Y	1	-0.3955	0.0066	-60.1292	0.0001	1.0000	1.0000

Fig 2:

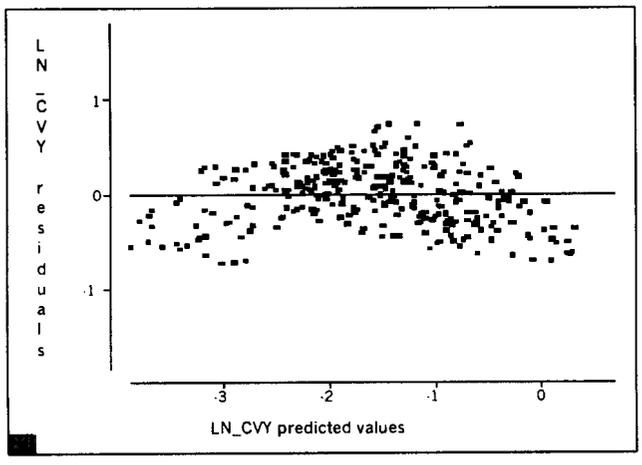
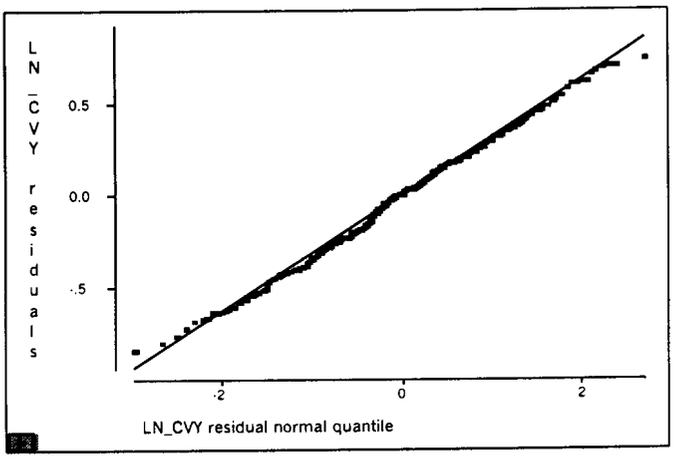


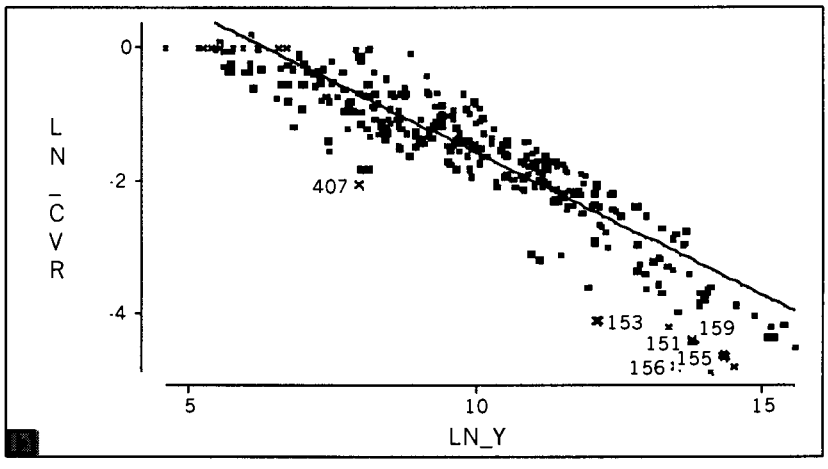
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different dwelling-types, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated population number of households with different dwelling-types. (Source: OHS 97 – Household file)

Model: $\ln(cv(\hat{R})) = 2.7167 - 0.4297 \ln(\hat{Y}_c)$

Fig: 1



Outliers have been excluded from calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
———	1	1	378.7764	429	0.1511	0.8539	2507.1317	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.7167	0.0872	31.1548	0.0001	.	0
LN Y	1	-0.4297	0.0086	-50.0713	0.0001	1.0000	1.0000

Fig 2:

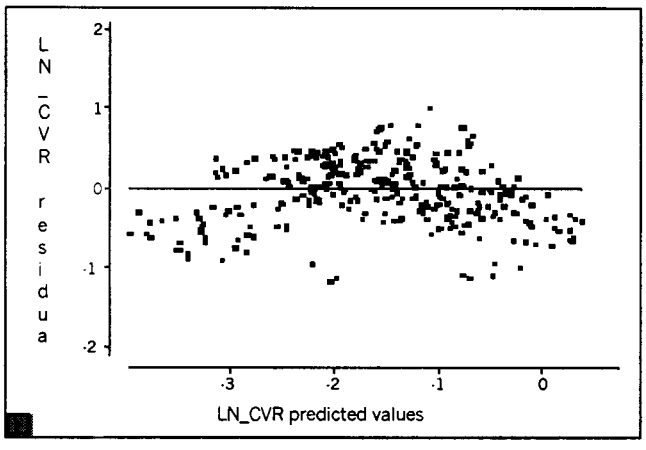
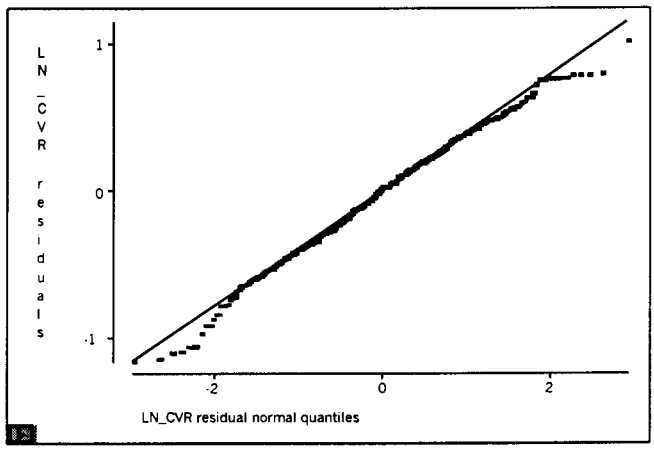


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa according to different sanitation facilities, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 97 - Household)

Model: $\ln(cv(\hat{Y}_c)) = 2.5463 - 0.415\ln(\hat{Y}_c)$

Fig: 1

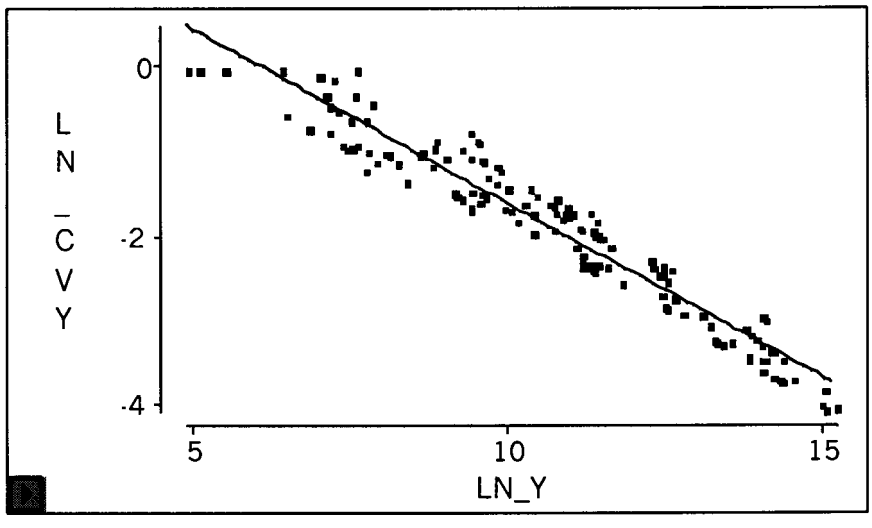


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	155.2889	153	0.0707	0.9349	2196.5834	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5463	0.0977	26.0699	0.0001	.	0
LN Y	1	-0.4150	0.0089	-46.8677	0.0001	1.0000	1.0000

Fig 2:

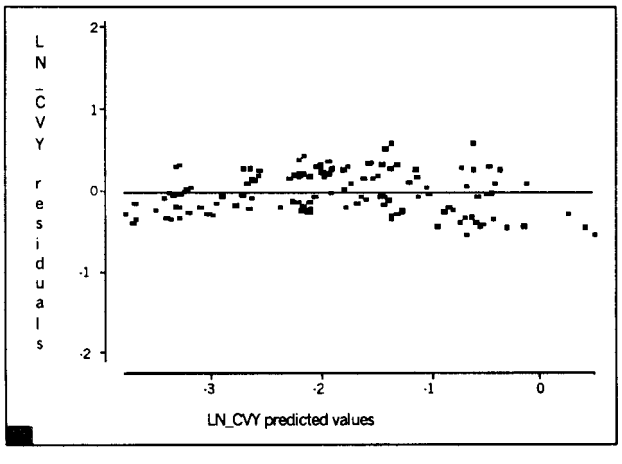
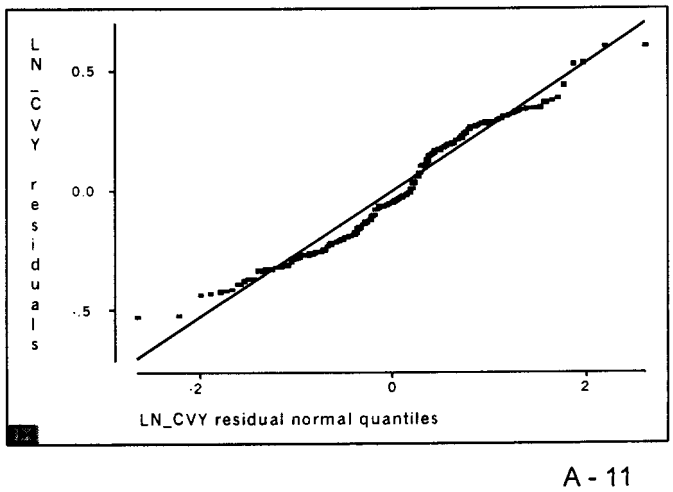


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different sanitation facilities, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated number of households with different sanitation facilities. (Source: OHS 97 – Household file)

Model: $\ln(cv(\hat{R})) = 2.6965 - 0.4368 \ln(\hat{Y}_c)$

Fig: 1

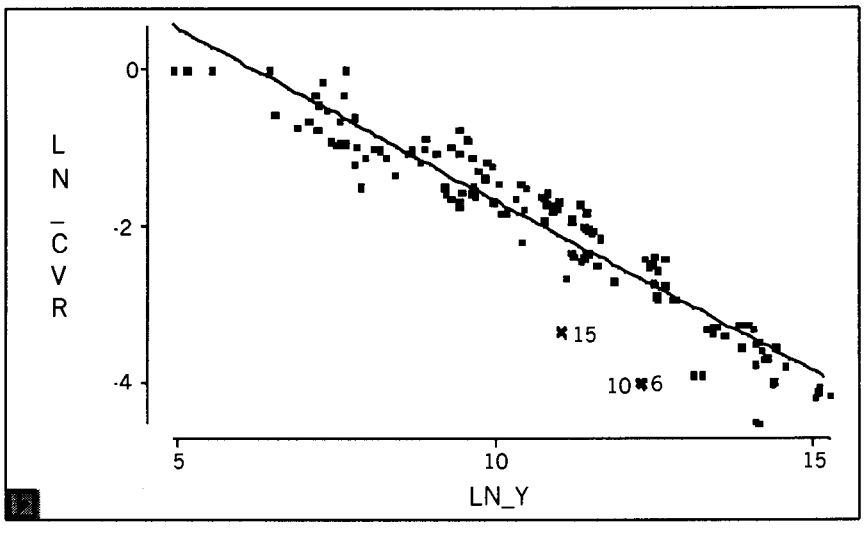


Table 1:

Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob>F
	1	1	171.0834	150	0.1072	0.9141	1595.5711	0.0001

Table 2:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.6965	0.1204	22.3911	0.0001	.	0
LN_Y	1	-0.4368	0.0109	-39.9446	0.0001	1.0000	1.0000

Fig 2:

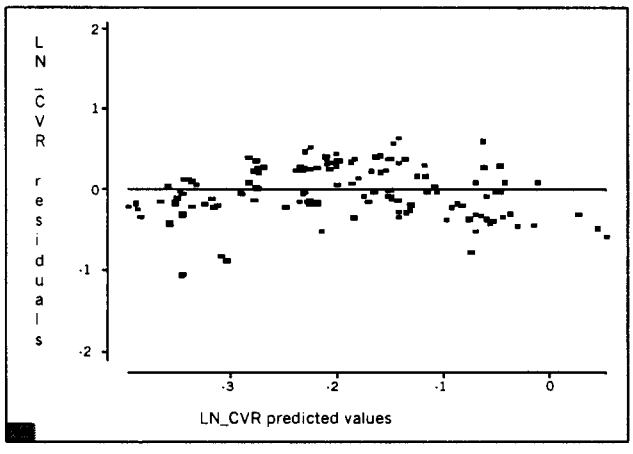
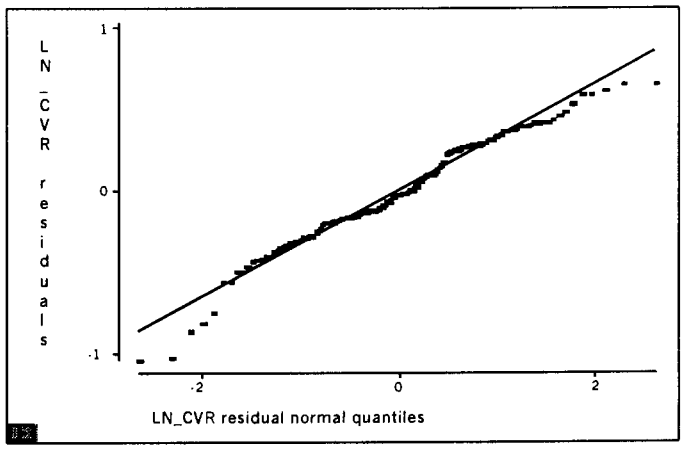


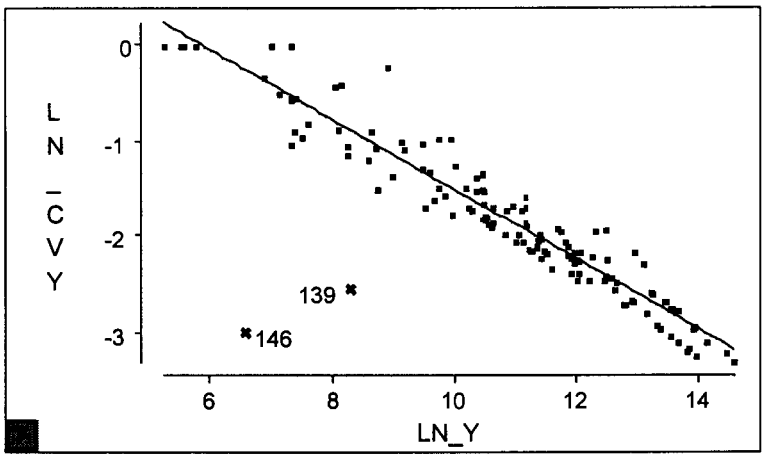
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population total unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 96 - Workers)

Model: $\ln(cv(\hat{Y}_c)) = 2.1526 - 0.364\ln(\hat{Y}_c)$

Fig 1:



Observations 139 and 146 are outliers and have been excluded from all calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	82.2852	139	0.0455	0.9286	1807.0452	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.1526	0.0954	22.5598	0.0001	.	0
LN_Y	1	-0.3649	0.0086	-42.5094	0.0001	1.0000	1.0000

Fig 2:

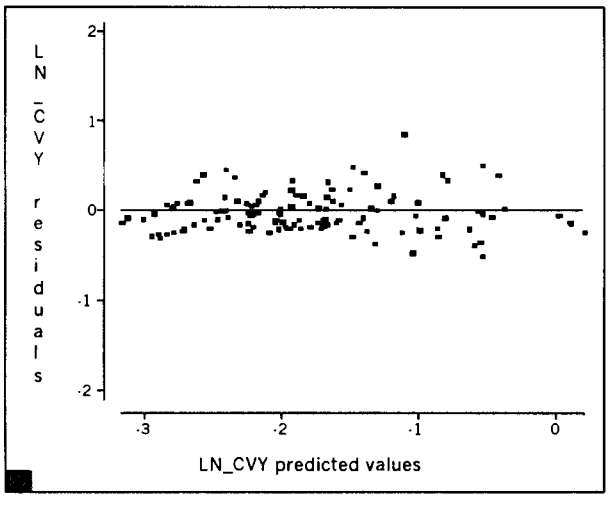
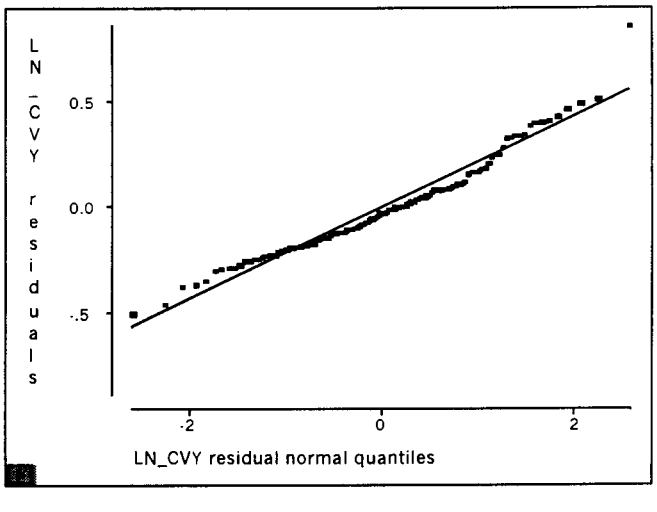


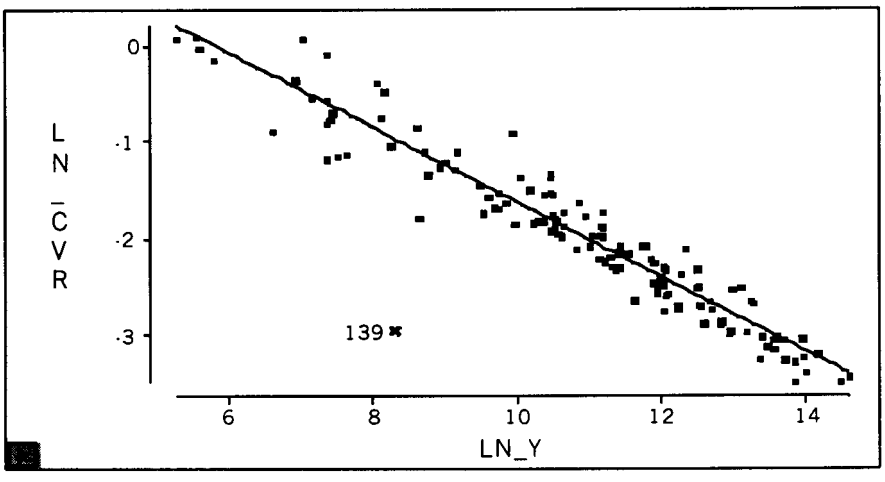
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 96 - Workers)

$$\textbf{Model: } \ln(cv(\hat{R})) = 2.2642 - 0.3879\ln(\hat{Y}_c)$$

Fig 1:



Observation 139 is excluded from the calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	95.7933	140	0.0510	0.9307	1879.6866	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.2642	0.0992	22.8162	0.0001	.	0
LN Y	1	-0.3879	0.0089	-43.3554	0.0001	1.0000	1.0000

Fig 2:

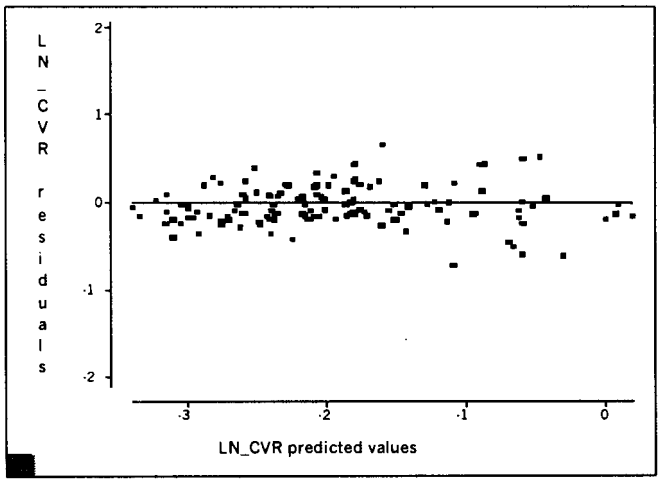
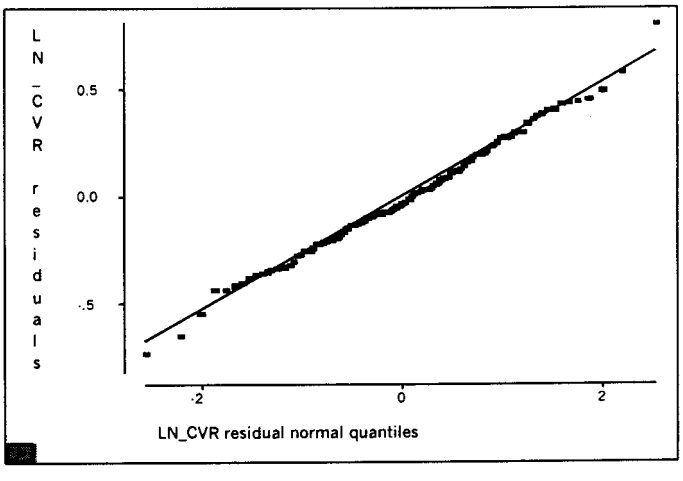


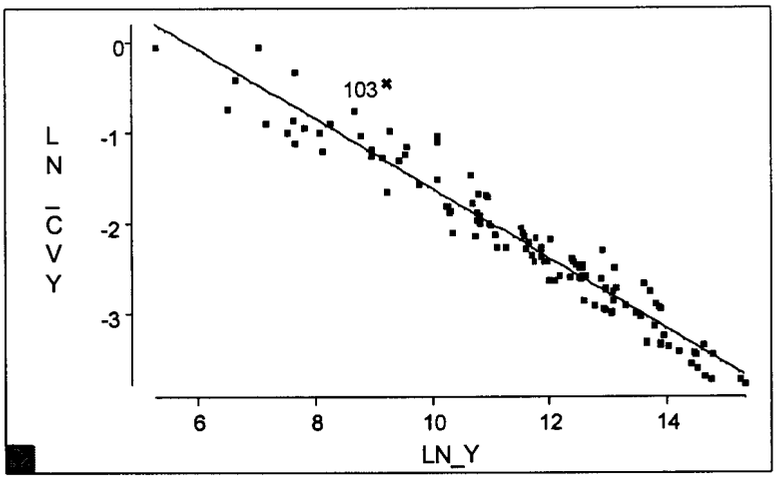
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population total unemployed in South Africa, according to the expanded definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$. (Source: OHS 96 - Workers)

Model: $\ln(cv(\hat{Y}_c)) = 2.2816 - 0.3877 \ln(\hat{Y}_c)$

Fig 1:



Observation 103 is excluded as an outlier.

Table 1:

Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
	1	1	82.3866	117	0.0434	0.9419	1896.9011	0.0001

Table 2:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.2816	0.1046	21.8103	0.0001	.	0
LN_Y	1	-0.3877	0.0089	-43.5534	0.0001	1.0000	1.0000

Fig 2:

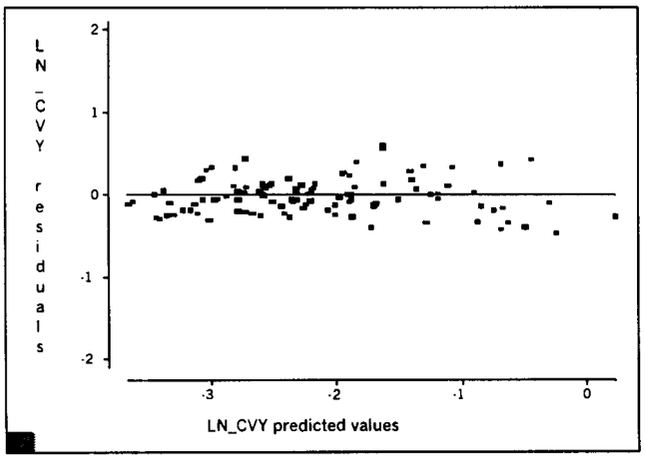
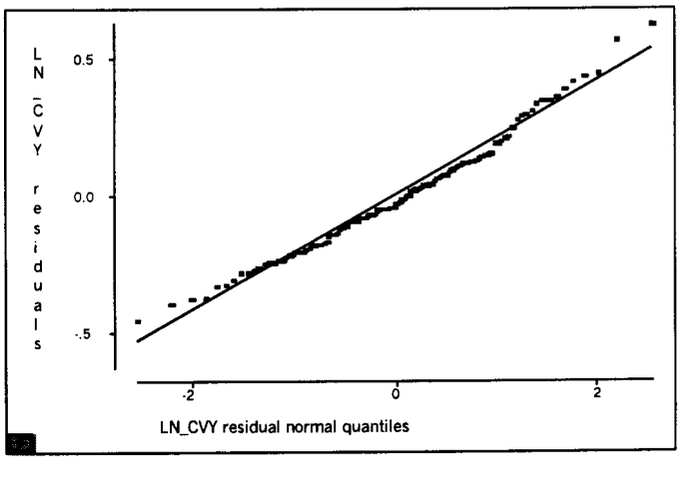


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio unemployed in South Africa, according to the expanded definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 96 - Workers)

$$\textbf{Model: } \ln(cv(\hat{R})) = 2.4506 - 0.4194\ln(\hat{Y}_c)$$

Fig 1:

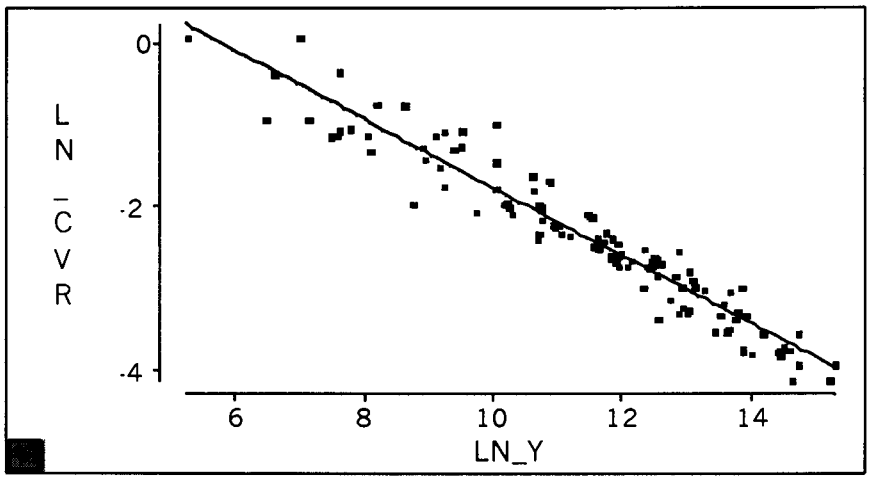


Table 1:

Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
———	1	1	97.3193	118	0.0701	0.9216	1387.6841	0.0001

Table 2:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.4506	0.1321	18.5539	0.0001	.	0
LN_Y	1	-0.4194	0.0113	-37.2516	0.0001	1.0000	1.0000

Fig 2:

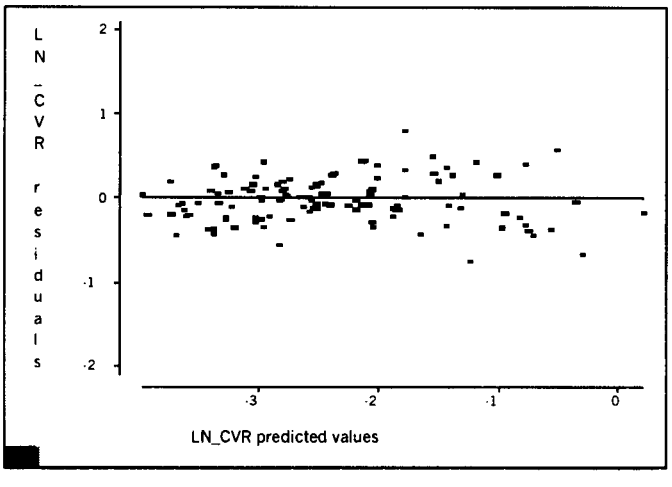
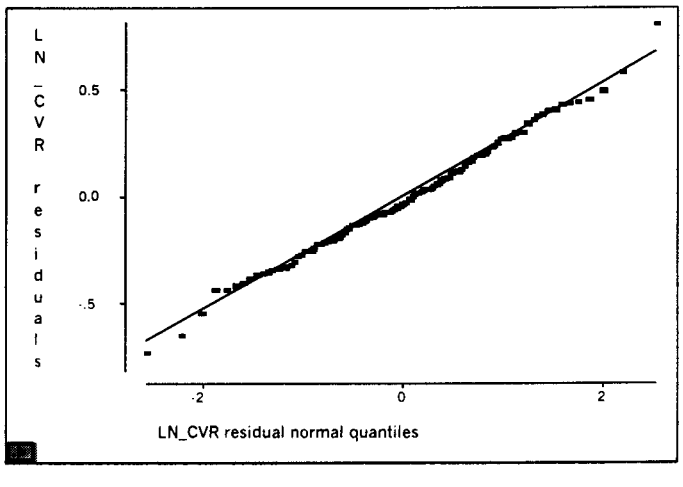


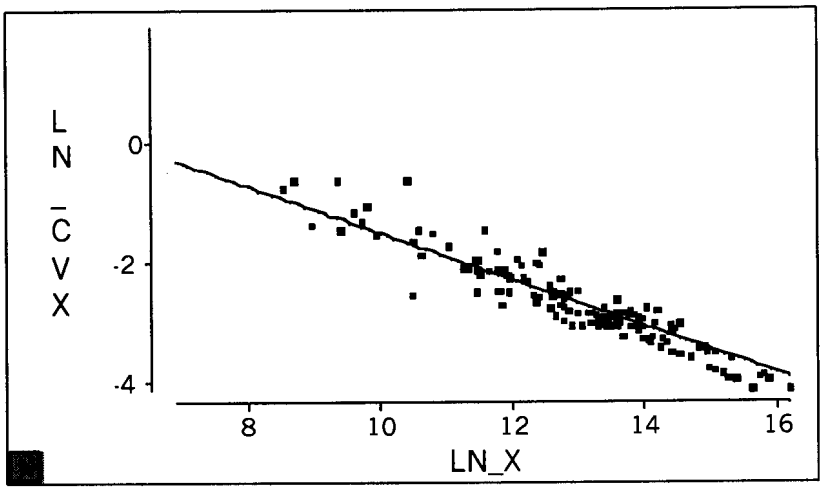
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of economic active people in South Africa as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 96 – Workers File)

Model: $\ln(cv(\hat{Y}_c)) = 2.9642 - 0.4347 \ln(\hat{Y}_c)$

Fig: 1



All outliers have been excluded from the calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	DF Model	Mean Square	DF Error	Mean Square	R-Square	F Stat	Prob > F
	1	1	65.9085	139	0.0565	0.8936	1165.5866	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob > |T|	Tolerance	Var Inflation
INTERCEPT	1	2.9642	0.1668	17.7679	0.0001	.	0
LN X	1	-0.4347	0.0127	-34.1407	0.0001	1.0000	1.0000

Fig 2:

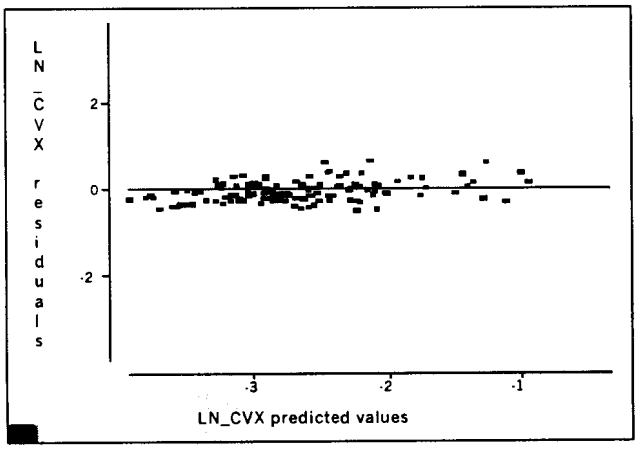
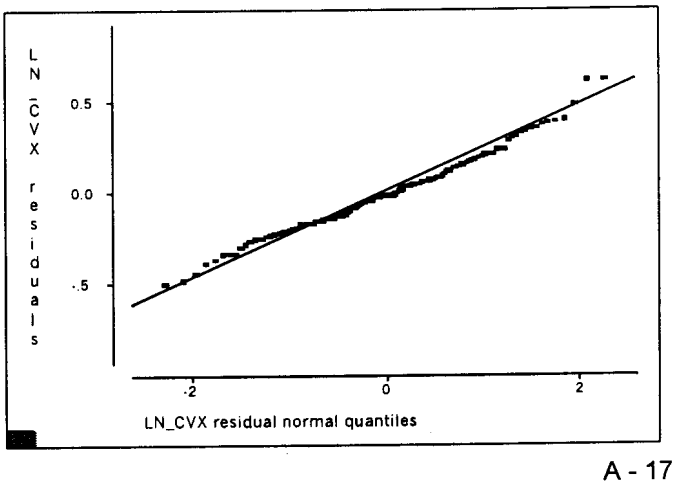


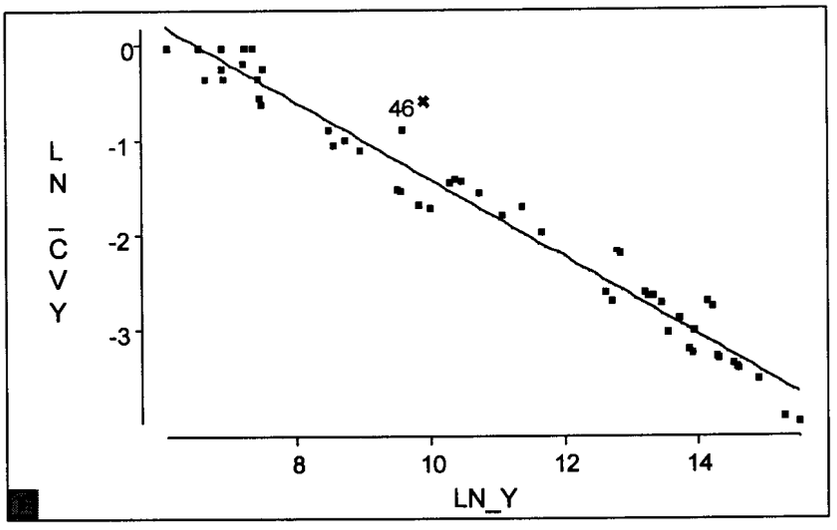
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa with different lighting-sources, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 96 – Household)

Model: $\ln(cv(\hat{Y}_c)) = 2.6678 - 0.4072\ln(\hat{Y}_c)$

Fig 1:



Observation 46 is excluded from calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	80.5491	53	0.0470	0.9700	1714.8839	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.6678	0.1109	24.0529	0.0001	.	0
LN_Y	1	-0.4072	0.0098	-41.4112	0.0001	1.0000	1.0000

Fig 2:

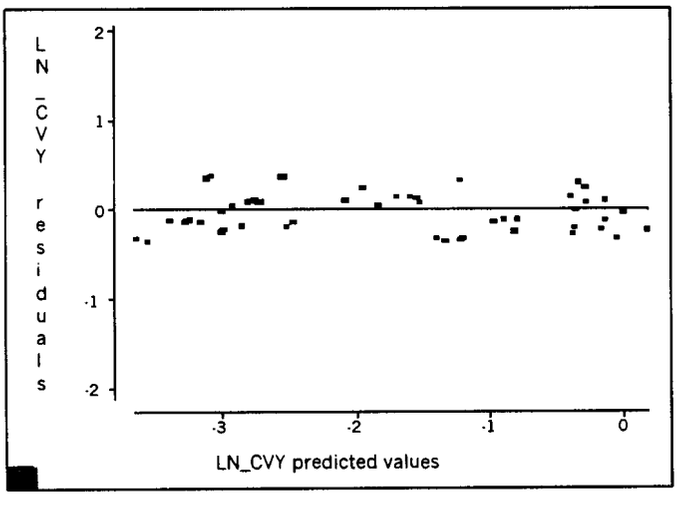
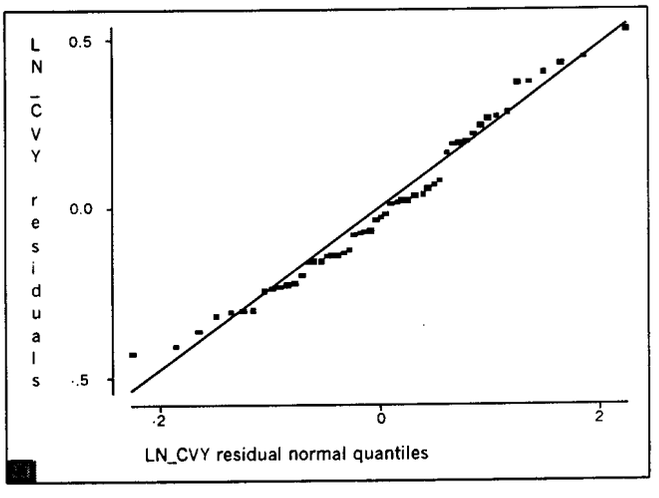


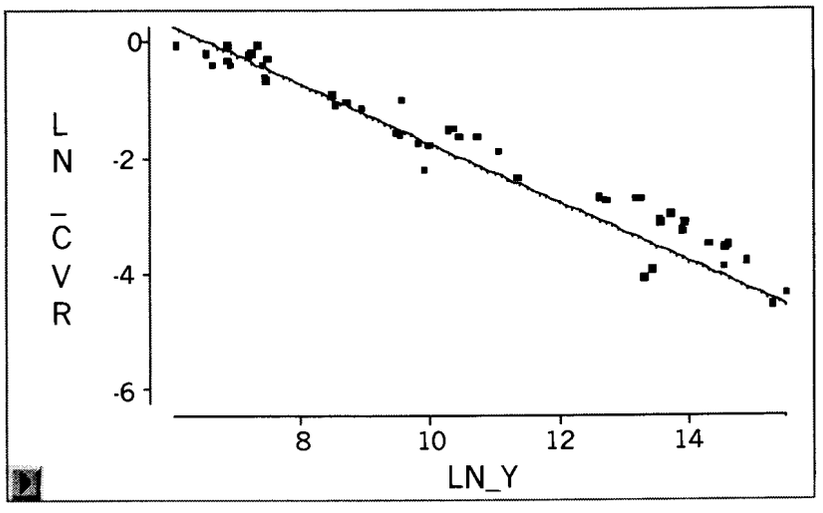
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different light sources, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated population number of households with different light sources.
(Source: OHS – 96 Household)

$$\textbf{Model: } \ln(cv(\hat{R})) = 2.9872 - 0.4563\ln(\hat{Y}_c)$$

Fig: 1



Observations identified as outliers have been excluded.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	95.4490	50	0.1605	0.9224	594.5374	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob > |T|	Tolerance	Var Inflation
INTERCEPT	1	2.9872	0.2075	14.3955	0.0001	.	0
LN_Y	1	-0.4563	0.0187	-24.3831	0.0001	1.0000	1.0000

Fig 2:

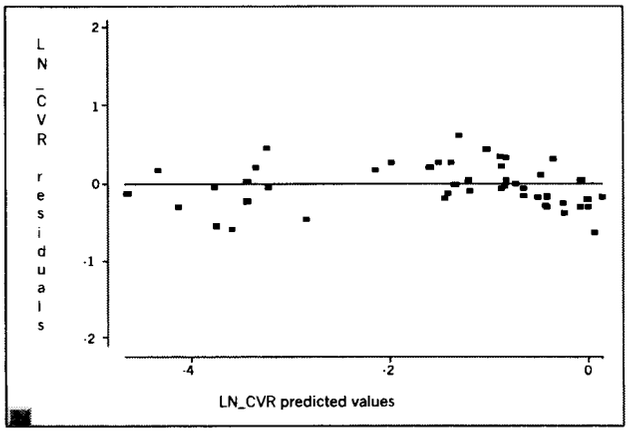
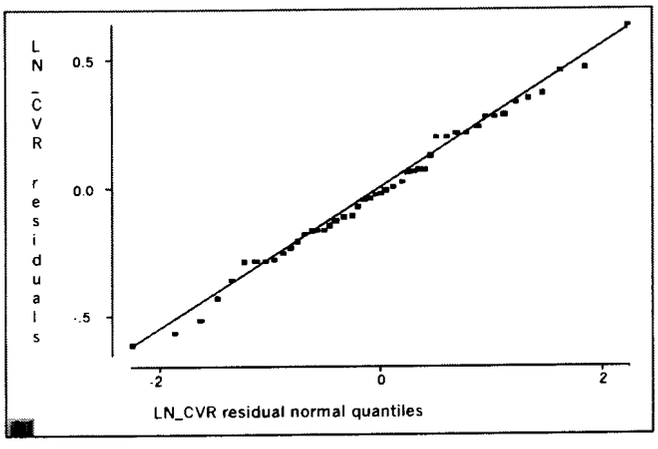


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa according to different water sources, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 96 – Household)

Model: $\ln(cv(\hat{Y}_c)) = 2.949 - 0.4154 \ln(\hat{Y}_c)$

Fig: 1

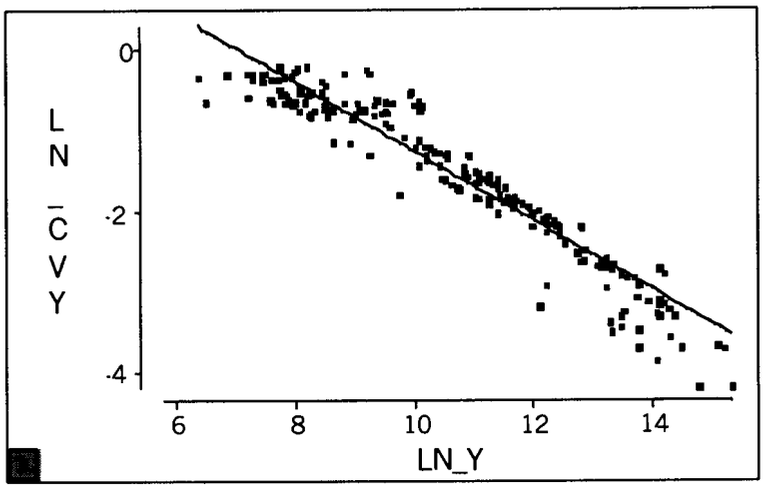


Table 1:

Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
	1	1	238.3185	300	0.0735	0.9153	3243.7239	0.0001

Table 2:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.9490	0.0774	38.1043	0.0001	.	0
LN_Y	1	-0.4154	0.0073	-56.9537	0.0001	1.0000	1.0000

Fig 2:

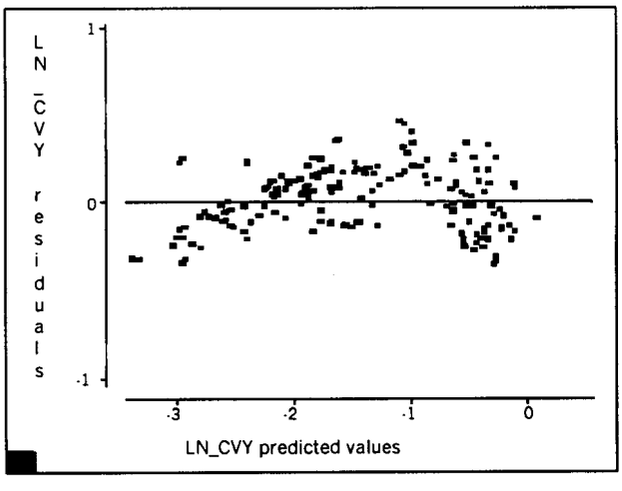
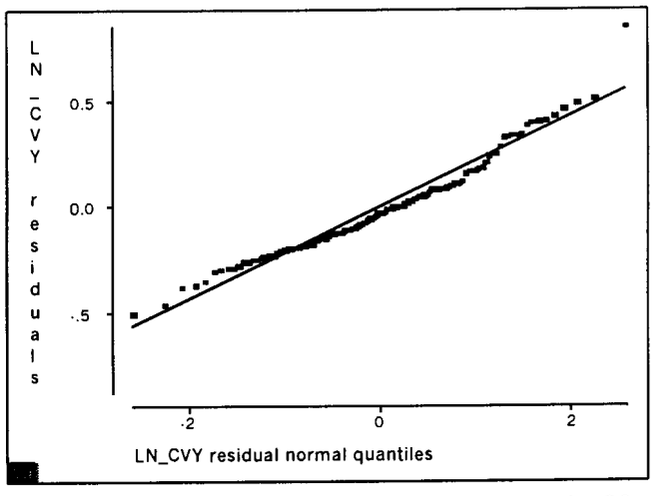


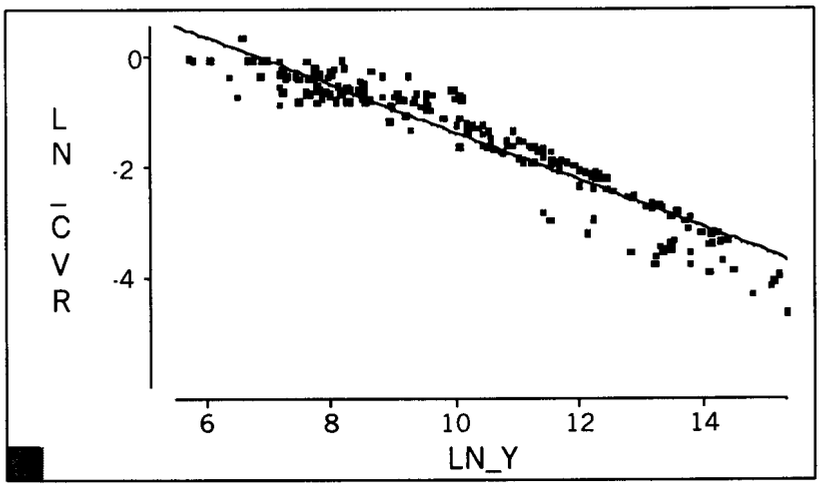
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different water sources, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated population number of households with different water sources.
(Source: OHS 96 - Household)

$$\textbf{Model: } \ln(cv(\hat{R})) = 2.9383 - 0.4227\ln(\hat{Y}_c)$$

Fig: 1



Observations identified as outliers have been excluded.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	350.9724	344	0.1155	0.8983	3037.7992	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.9383	0.0786	37.3625	0.0001	.	0
LNY	1	-0.4227	0.0077	-55.1162	0.0001	1.0000	1.0000

Fig 2:

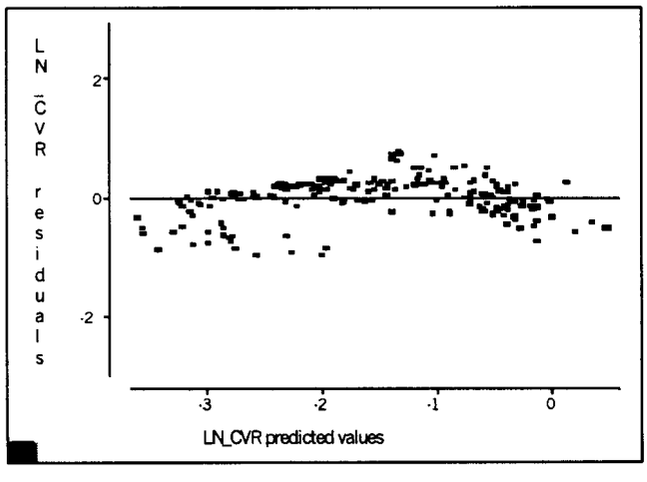
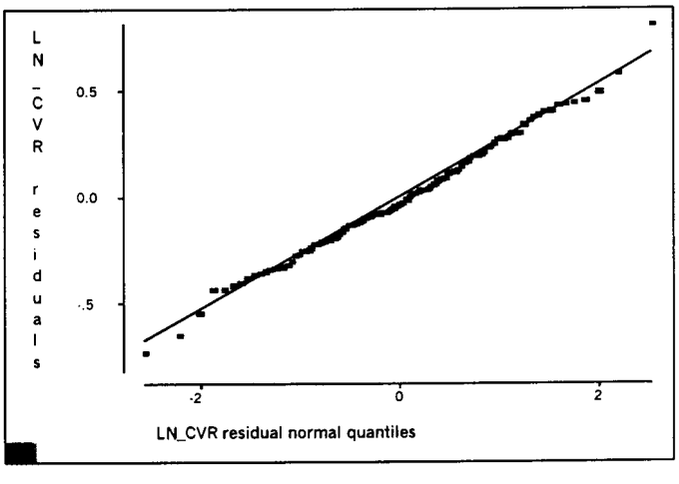


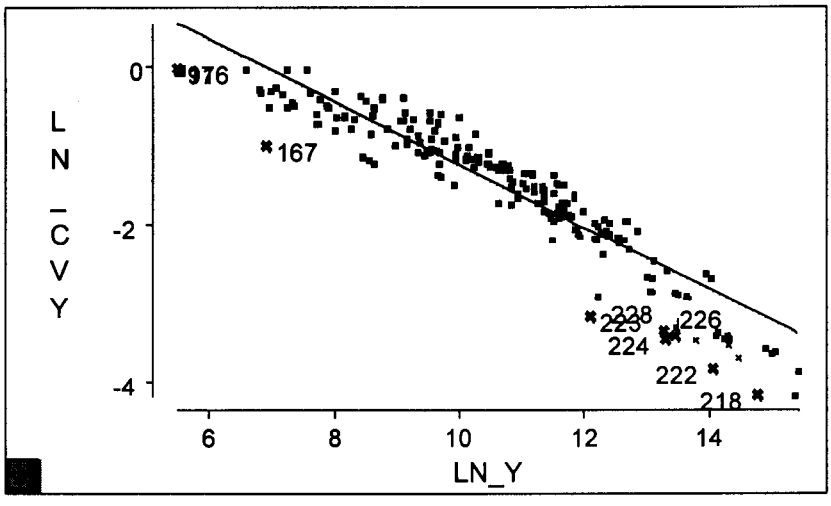
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa according to different dwelling-types, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 96 - Household)

$$\textbf{Model: } \ln(cv(\hat{Y}_c)) = 2.7282 - 0.3957\ln(\hat{Y}_c)$$

Fig 1:



All the observations marked with x have been identified as outliers and are excluded from the calculations

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	214.1395	291	0.0810	0.9008	2642.4421	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.7280	0.0818	33.3363	0.0001	.	0
LN_Y	1	-0.3957	0.0077	-51.4047	0.0001	1.0000	1.0000

Fig 2:

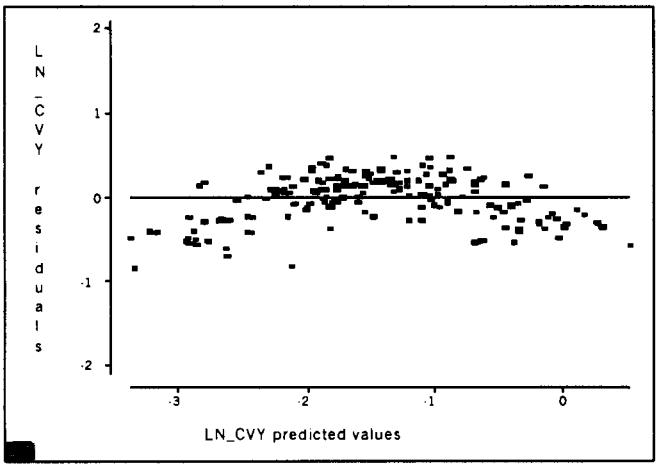
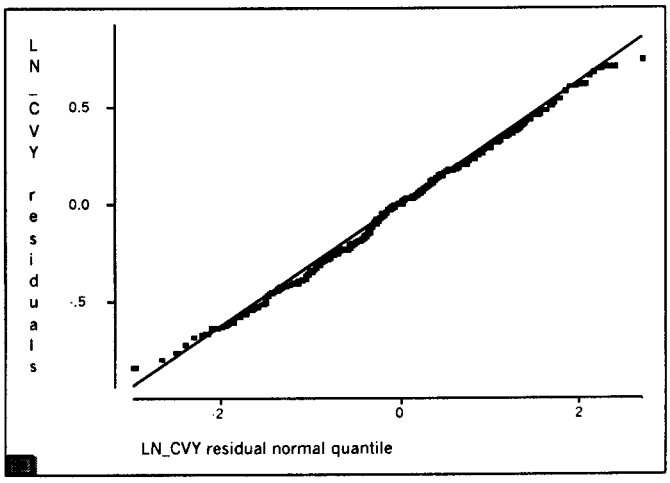


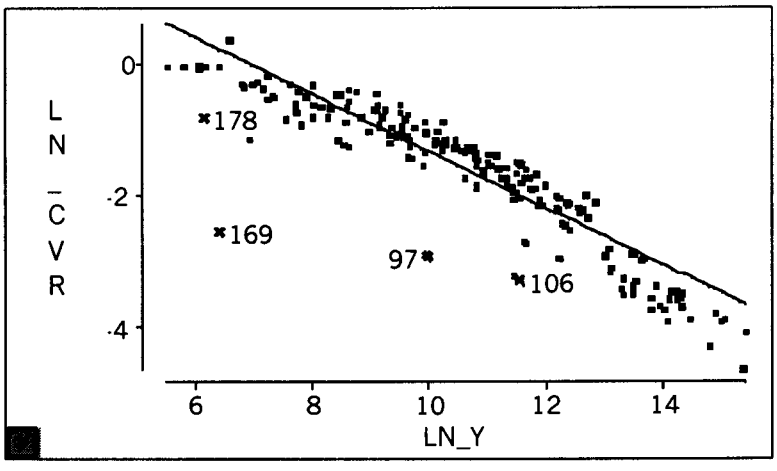
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different dwelling-types, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated population number of households with different dwelling-types.
(Source: OHS 96 – Household)

Model: $\ln(cv(\hat{R})) = 3.025 - 0.4325 \ln(\hat{Y}_c)$

Fig: 1



Outliers have been excluded from calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	284.2134	299	0.1245	0.8842	2281.9394	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	3.0250	0.0971	31.1541	0.0001	.	0
LN_Y	1	-0.4325	0.0091	-47.7696	0.0001	1.0000	1.0000

Fig 2:

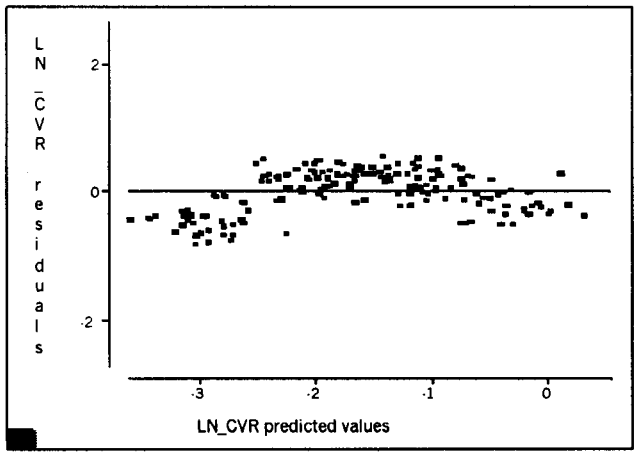
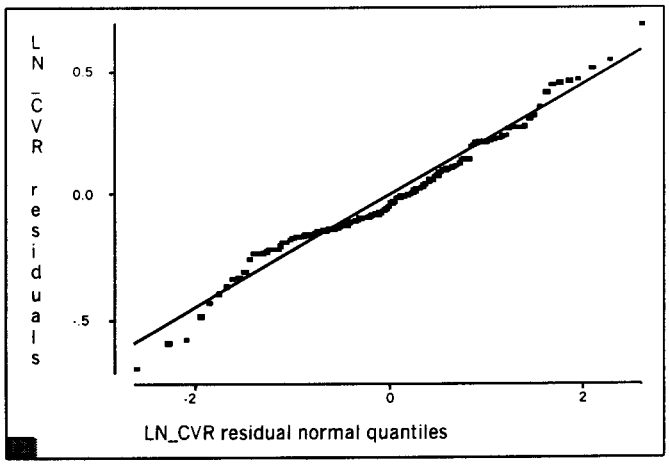


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa according to different sanitation facilities, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 96 - Household)

Model: $\ln(cv(\hat{Y}_c)) = 3.3437 - 0.4512\ln(\hat{Y}_c)$

Fig 1:

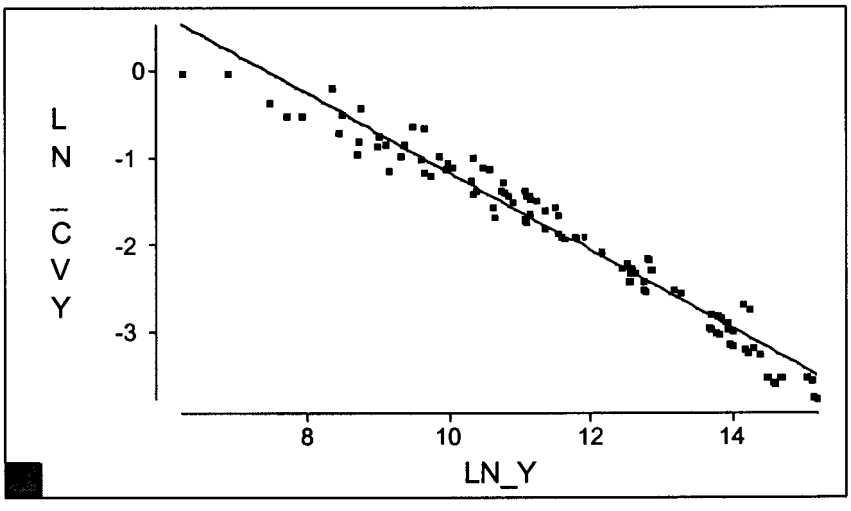


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	97.5750	107	0.0385	0.9595	2533.1428	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	3.3437	0.1055	31.7054	0.0001	.	0
LN_Y	1	-0.4512	0.0090	-50.3303	0.0001	1.0000	1.0000

Fig 2:

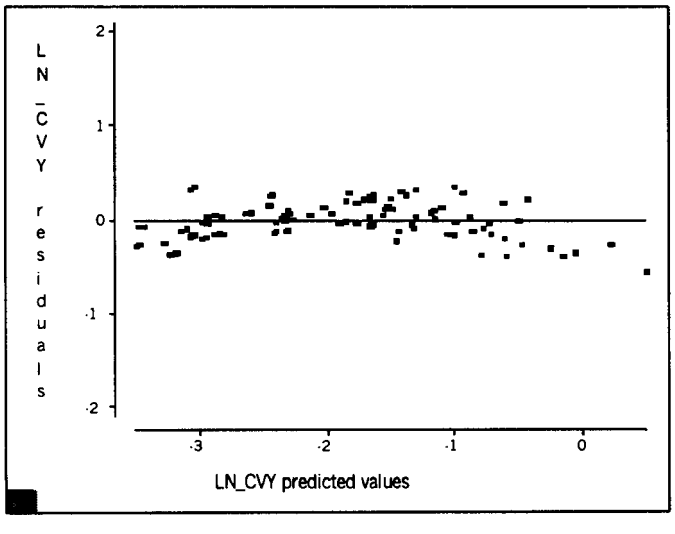
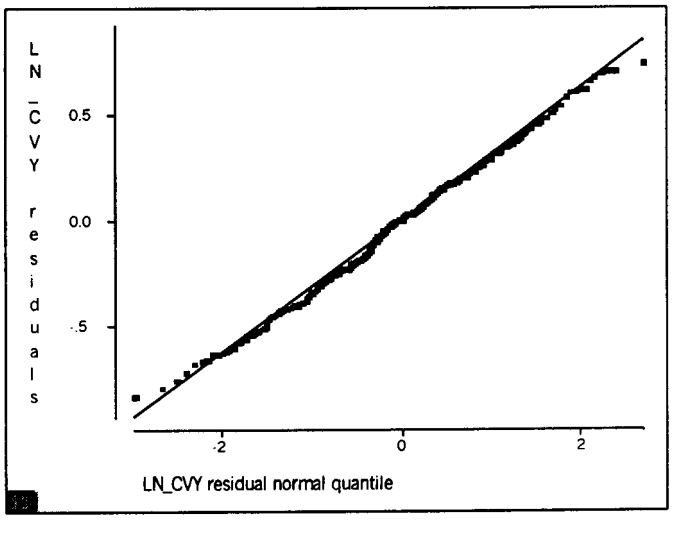


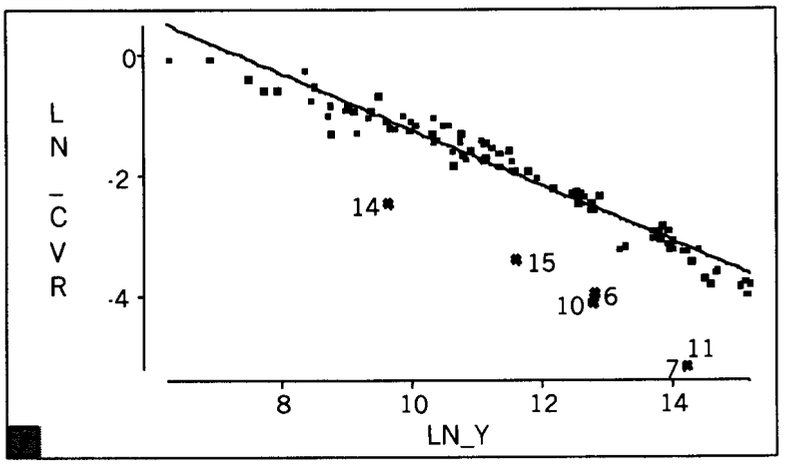
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different sanitation facilities, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated population number of households with different sanitation facilities.
(Source: OHS 96 - Household)

$$\textbf{Model: } \ln(cv(\hat{R})) = 3.4542 - 0.4671\ln(\hat{Y}_c)$$

Fig: 1



All the outliers have been excluded from the calculations.

Table 1:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	3.4542	0.1237	27.9327	0.0001	.	0
LN Y	1	-0.4671	0.0106	-44.2328	0.0001	1.0000	1.0000

Table 2:

Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob>F
	1	1	100.0388	101	0.0511	0.9509	1956.5431	0.0001

Fig 2:

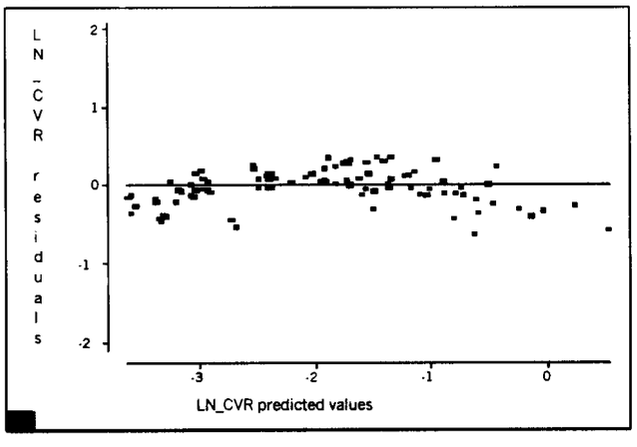
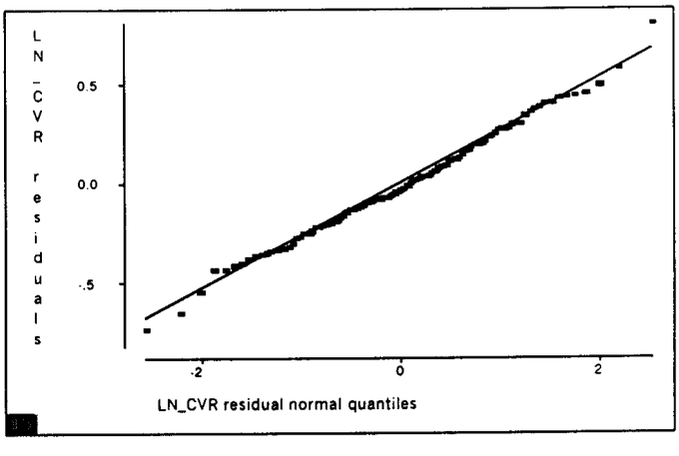


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 95 – Workers)

$$\textbf{Model: } \ln(cv(\hat{Y}_c)) = 1.1842 - 0.3895 \ln(\hat{Y}_c)$$

Fig 1:

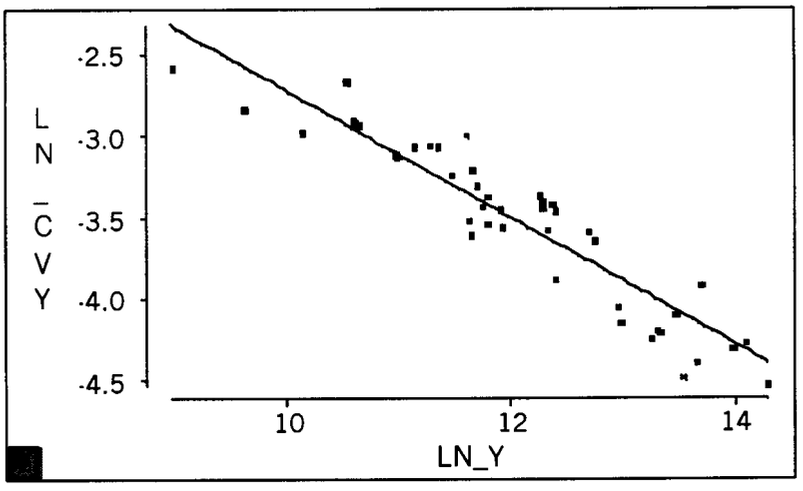


Table 1:

Parametric Regression								
			Model		Error			
Curve	Degree(Polynomia	DF	Mean	DF	Mean	R-	F Stat	Prob > F
	1	1		40		0.8890		

Table 2:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	1.1842	0.2632	4.4998	0.0001	.	0
LN_Y	1	-0.3895	0.0218	-17.9024	0.0001	1.0000	1.0000

Fig 2:

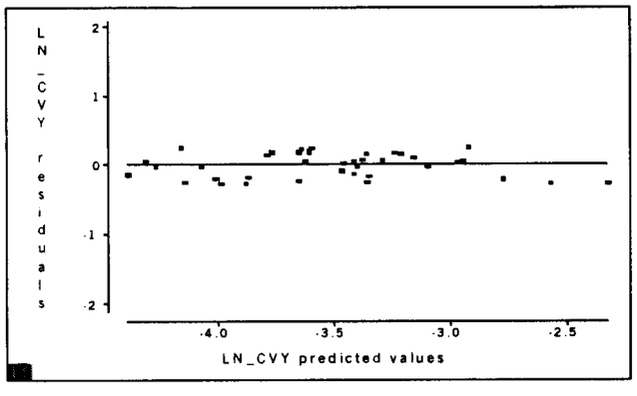
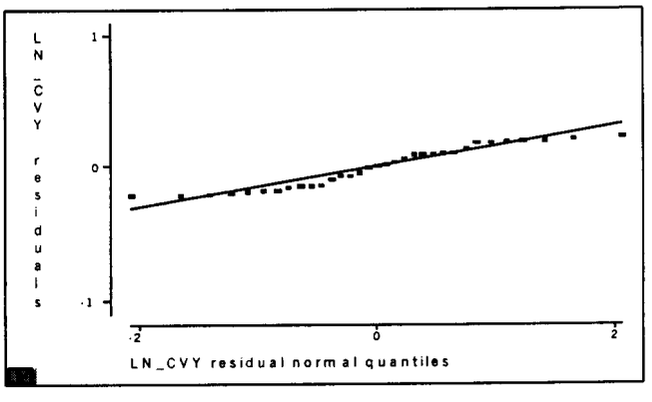


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population rate of unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 95 - Workers)

Model: $\ln(cv(\hat{R})) = 2.1541 - 0.4099\ln(\hat{Y}_c)$

Fig 1:

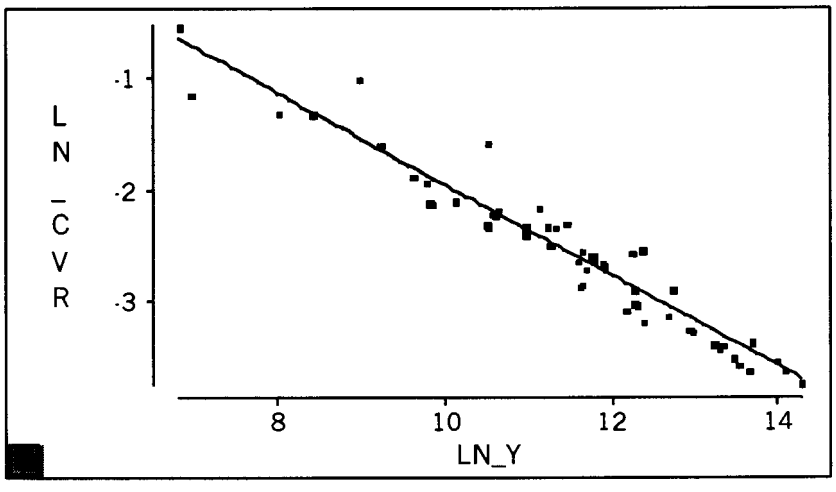


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	DF	Model Mean Square	DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	26.4140	54	0.0395	0.9252	667.9412	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.1541	0.1843	11.6855	0.0001	.	0
LN_Y	1	-0.4099	0.0159	-25.8446	0.0001	1.0000	1.0000

Fig 2:

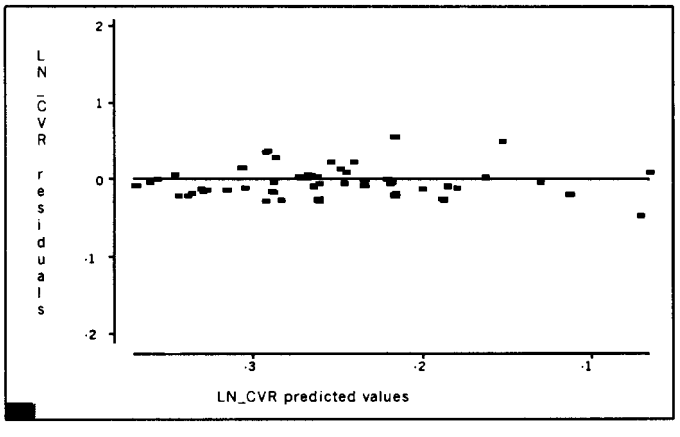
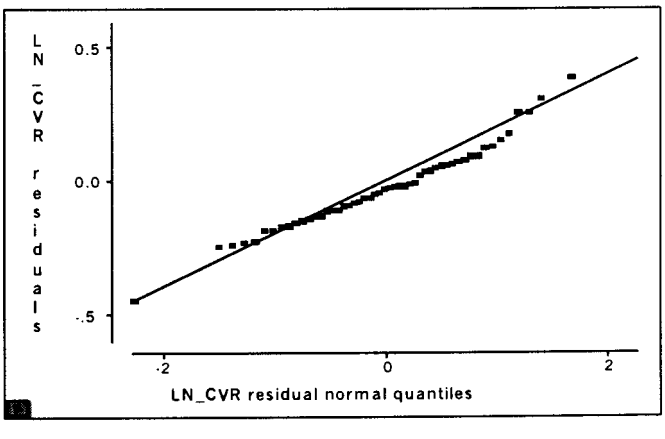


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa with different lighting-sources, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 95 - Household)

Model: $\ln(cv(\hat{Y}_c)) = 2.5066 - 0.4380\ln(\hat{Y}_c)$

Fig: 1

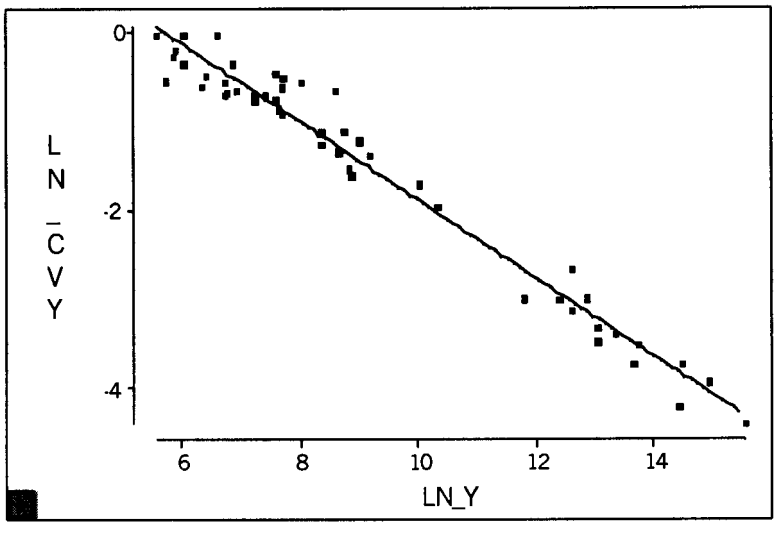


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	88.4429	51	0.0508	0.9715	1741.3000	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5066	0.1028	24.3772	0.0001	.	0
LN Y	1	-0.4380	0.0105	-41.7289	0.0001	1.0000	1.0000

Fig 2:

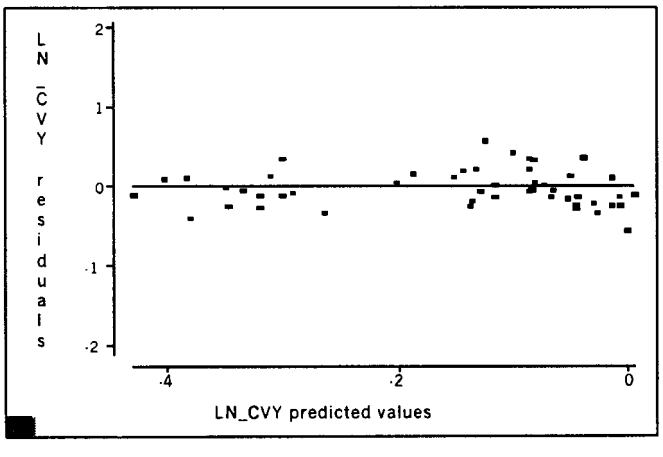
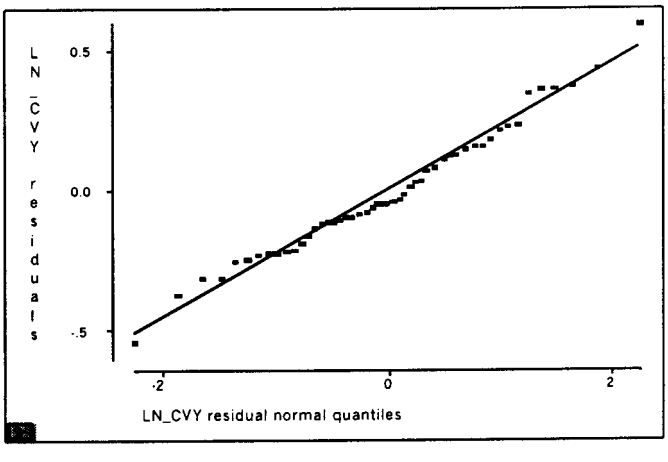


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different lighting-sources, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated population number of households with different lighting-sources.
(Source: OHS 95 - Household)

Model: $\ln(cv(\hat{R})) = 2.8218 - 0.4806\ln(\hat{Y}_c)$

Fig: 1

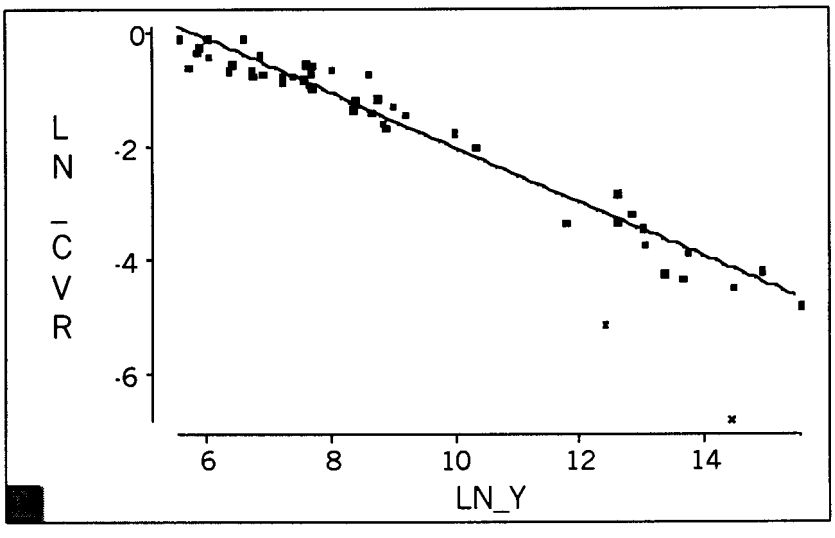


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	98.0512	49	0.0773	0.9628	1268.6378	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.8218	0.1299	21.7271	0.0001	.	0
LN Y	1	-0.4806	0.0135	-35.6179	0.0001	1.0000	1.0000

Fig 2:

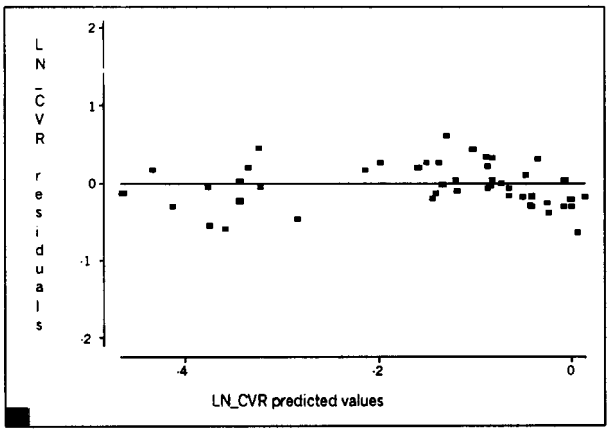
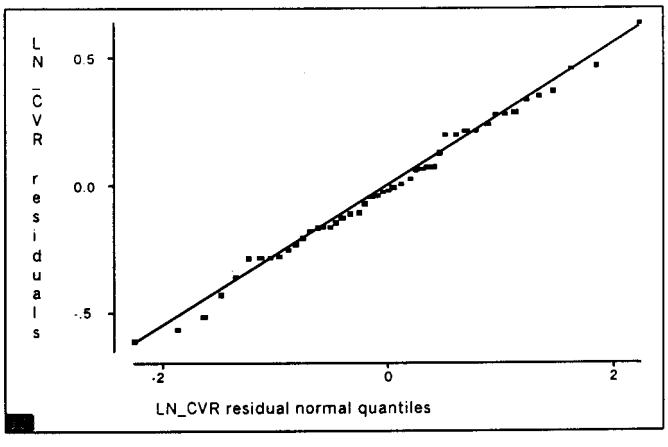


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa with different water sources, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 95 – Households)

$$\textbf{Model: } \ln(cv(\hat{Y}_c)) = 2.5541 - 0.4141\ln(\hat{Y}_c)$$

Fig: 1

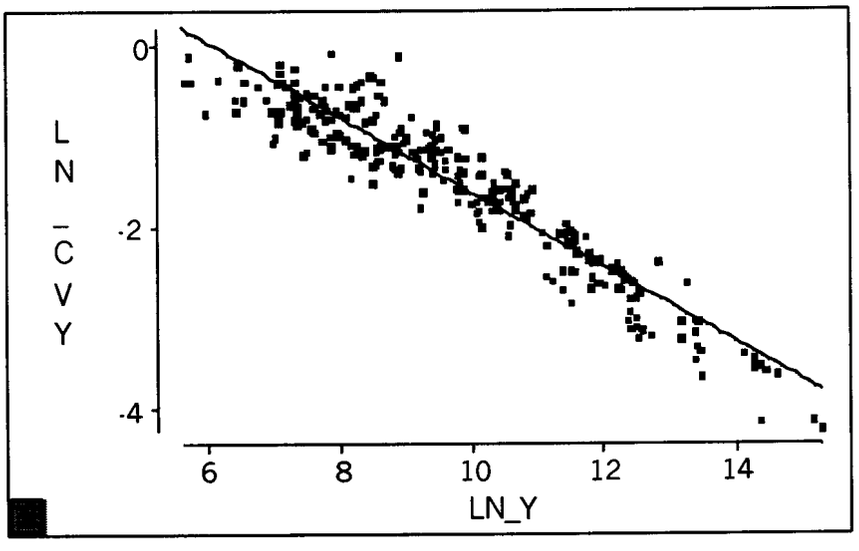


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
———	1	1	247.0396	338	0.0836	0.8974	2956.1532	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5541	0.0770	33.1762	0.0001	.	0
LN_Y	1	-0.4141	0.0076	-54.3705	0.0001	1.0000	1.0000

Fig 2:

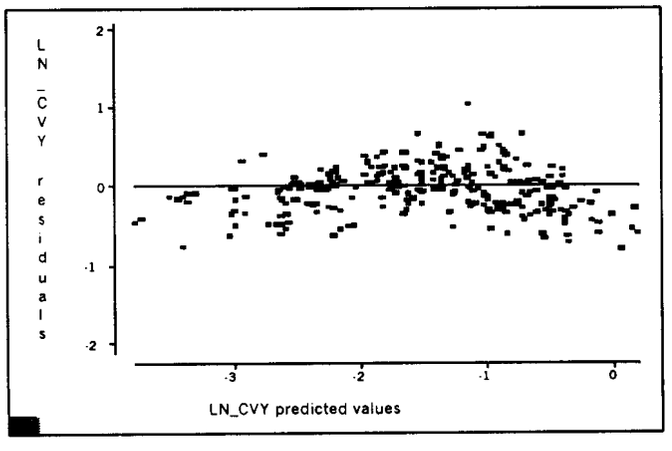
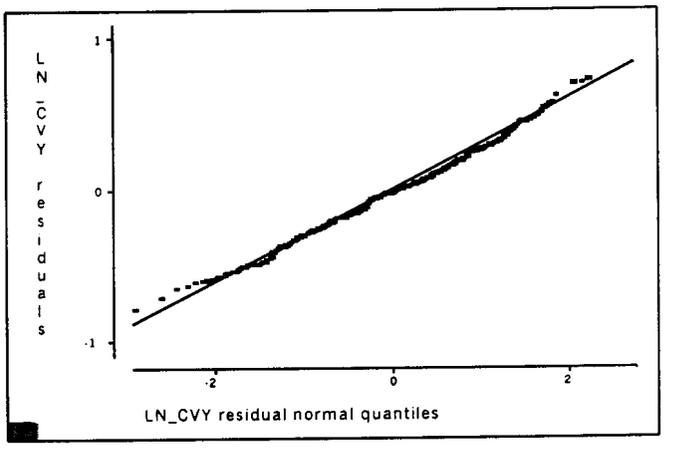


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different water sources, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated number of households with different water sources.
(Source: OHS 95 - Household)

Model: $\ln(cv(\hat{R})) = 2.4926 - 0.41\ln(\hat{Y}_c)$

Fig: 1

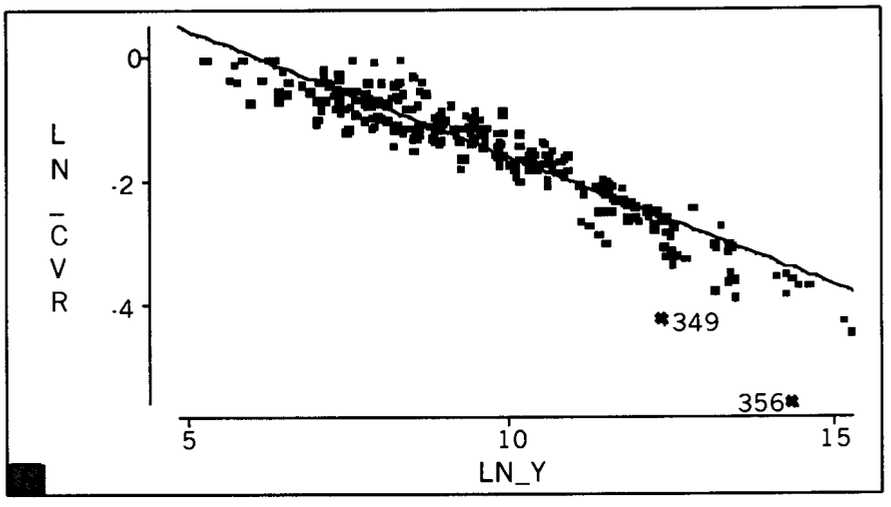


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	289.2557	353	0.0913	0.8998	3168.6511	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.4926	0.0722	34.5120	0.0001	.	0
LN_Y	1	-0.4100	0.0073	-56.2908	0.0001	1.0000	1.0000

Fig 2:

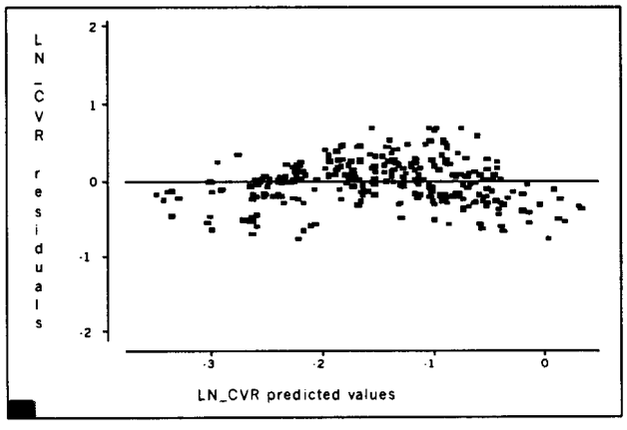
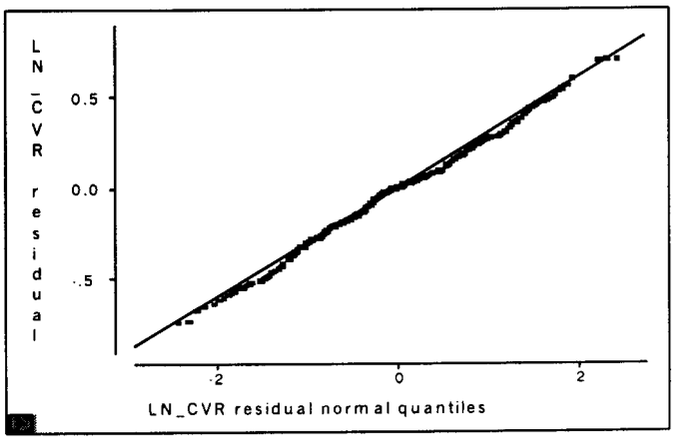


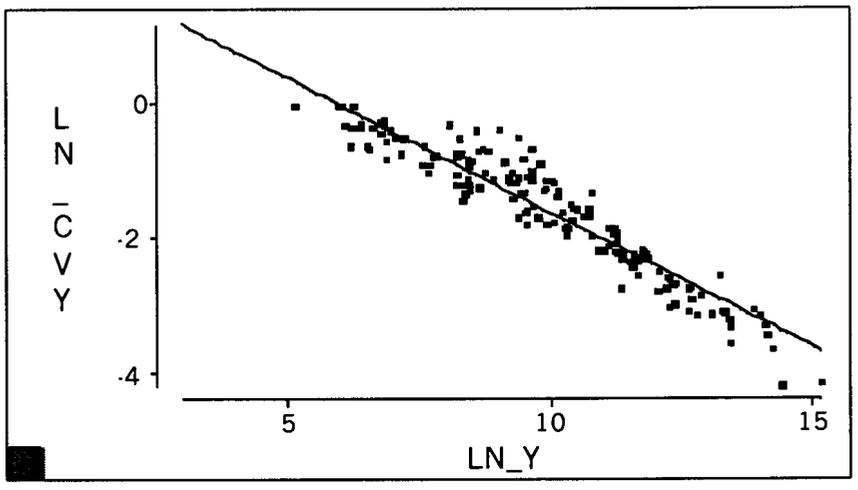
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated population number of households in South Africa with different sanitation facilities, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: OHS 95 - Households)

$$\textbf{Model: } \ln(cv(\hat{Y}_c)) = 2.3503 - 0.397\ln(\hat{Y}_c)$$

Fig: 1



Outliers were excluded from calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	167.9756	178	0.0931	0.9102	1804.1381	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob > |T|	Tolerance	Var Inflation
INTERCEPT	1	2.3503	0.0947	24.8075	0.0001	.	0
LN_Y	1	-0.3970	0.0093	-42.4751	0.0001	1.0000	1.0000

Fig 2:

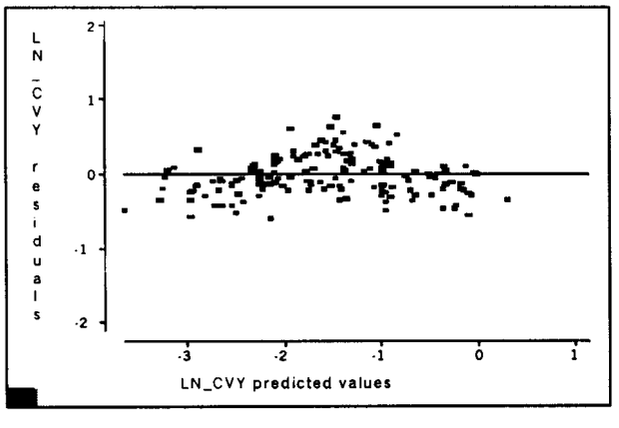
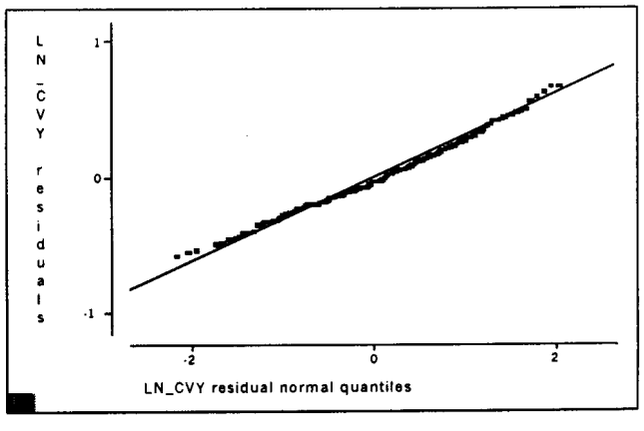


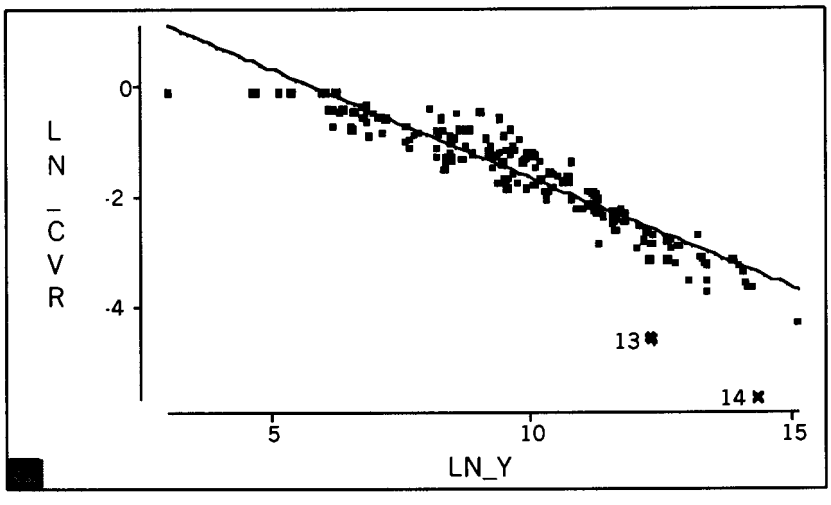
Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated population ratio of households in South Africa with different sanitation facilities, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated population number of households with different sanitation facilities.
(Source: OHS 95 - Households)

Model: $\ln(cv(\hat{R})) = 2.3707 - 0.3998\ln(\hat{Y}_c)$

Fig: 1



Outliers were excluded from calculations.

Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	165.9257	176	0.0974	0.9064	1703.4636	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob>|T|	Tolerance	Var Inflation
INTERCEPT	1	2.3707	0.0978	24.2484	0.0001	.	0
LN_Y	1	-0.3998	0.0097	-41.2730	0.0001	1.0000	1.0000

Fig 2:

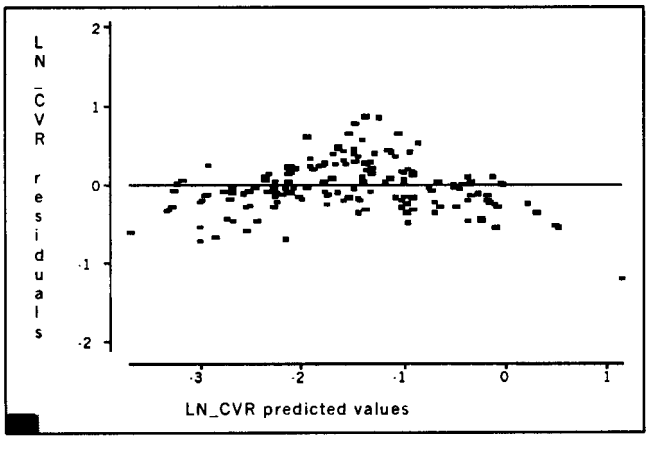
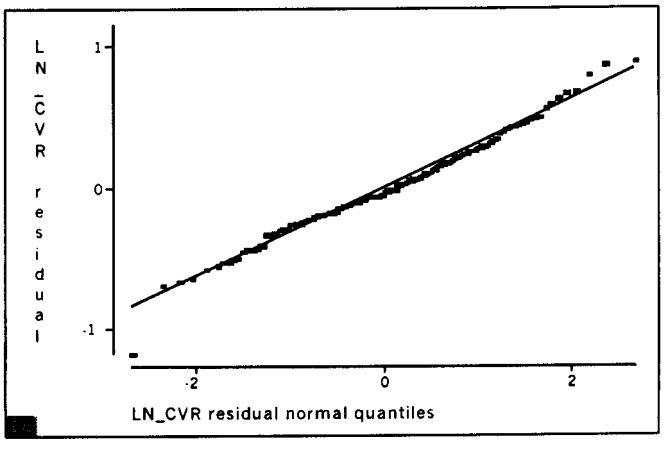


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated number of household crimes in South Africa, that includes all crimes that affect a household, as predicted by the natural logarithm of $\hat{Y}_c$. (Source: VOC -1998)

Model: $\ln(cv(\hat{Y}_c)) = 3.833 - 0.4436 \ln(\hat{Y}_c)$

Fig 1:

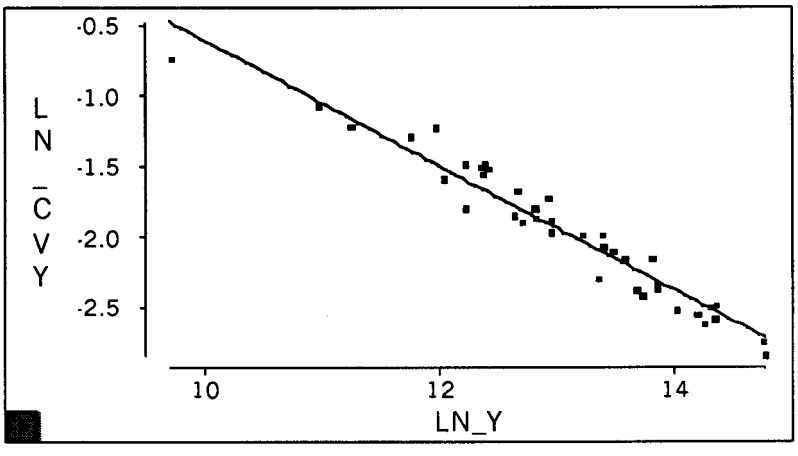


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	13.4234	60	0.0225	0.9086	596.1983	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	3.8330	0.2353	16.2918	0.0001	.	0
LN_Y	1	-0.4436	0.0182	-24.4172	0.0001	1.0000	1.0000

Fig 2:

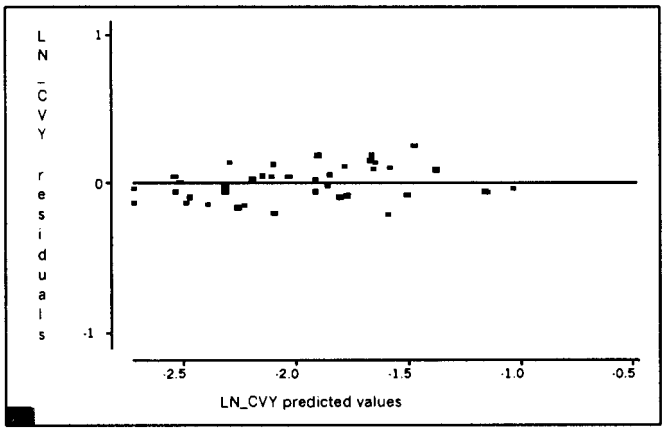
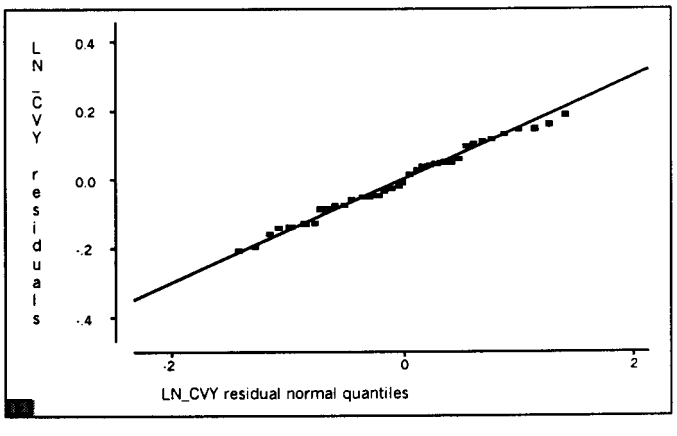


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{R}$, the estimated ratio of household crimes in South Africa, that includes all crimes that affect a household, as predicted by the natural logarithm of $\hat{Y}_c$, the estimated number of household crimes.
(Source: VOC -1998)

Model: $\ln(cv(\hat{R})) = 4.0889 - 0.4669\ln(\hat{Y}_c)$

Fig 1:

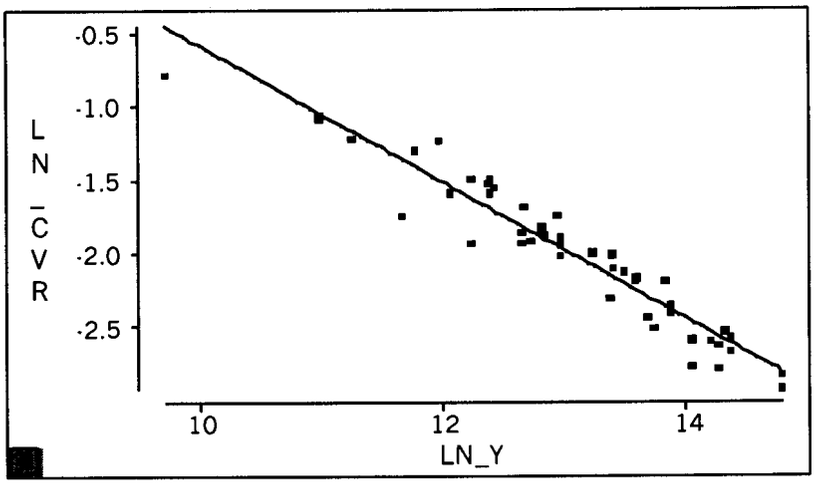


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	14.3782	58	0.0354	0.8751	406.2784	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	4.0889	0.2993	13.6636	0.0001	.	0
LN_Y	1	-0.4669	0.0232	-20.1563	0.0001	1.0000	1.0000

Fig 2:

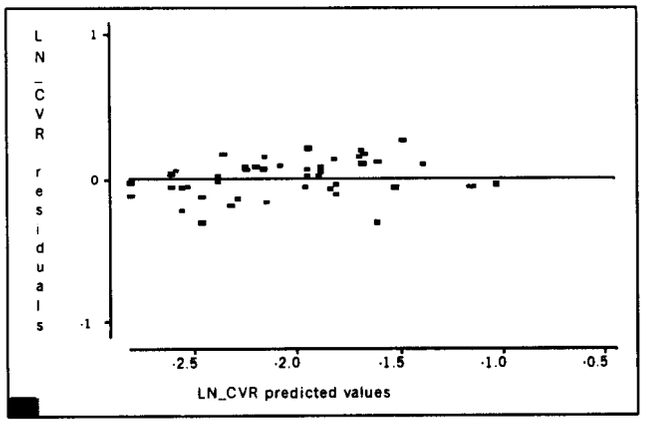
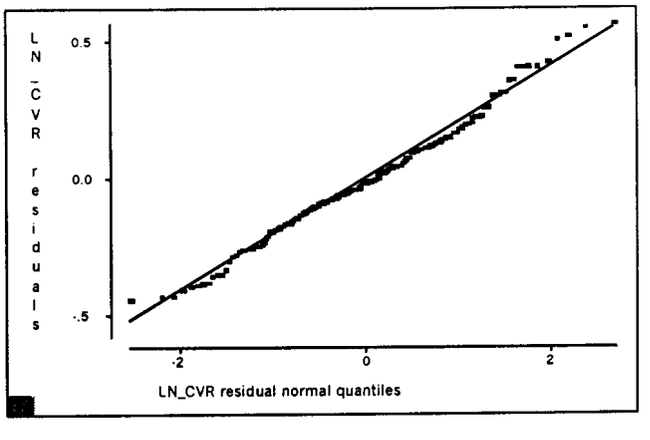


Fig 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}_c$, the estimated number of personal crimes in South Africa, that includes all crimes that affect an individual, as predicted by the natural logarithm of $\hat{Y}_c$.
(Source: VOC - 1998)

Model: $\ln(cv(\hat{Y}_c)) = 4.5238 - 0.498\ln(\hat{Y}_c)$

Fig 1:

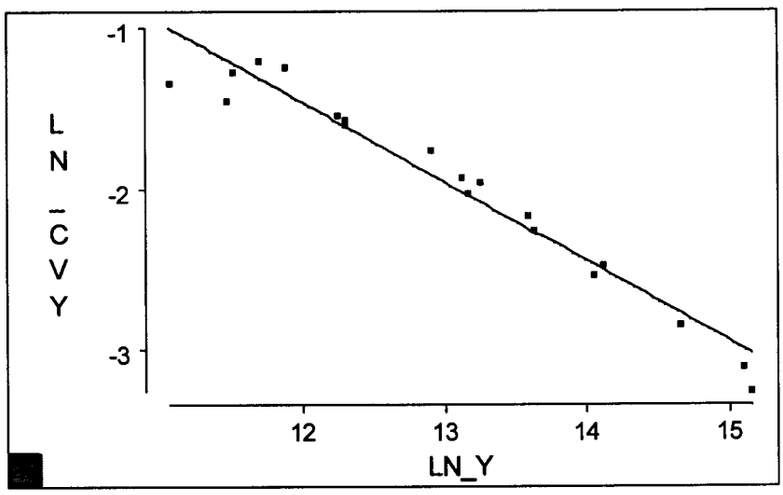


Table 1:

Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Prob > F
	1	1	7.1859	18	0.0195	0.9535	369.4092	0.0001

Table 2:

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	4.5238	0.3398	13.3132	0.0001	.	0
LN_Y	1	-0.4980	0.0259	-19.2200	0.0001	1.0000	1.0000

Fig 2:

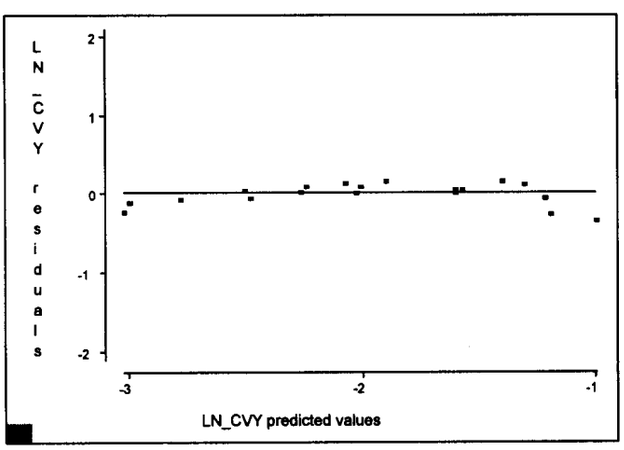
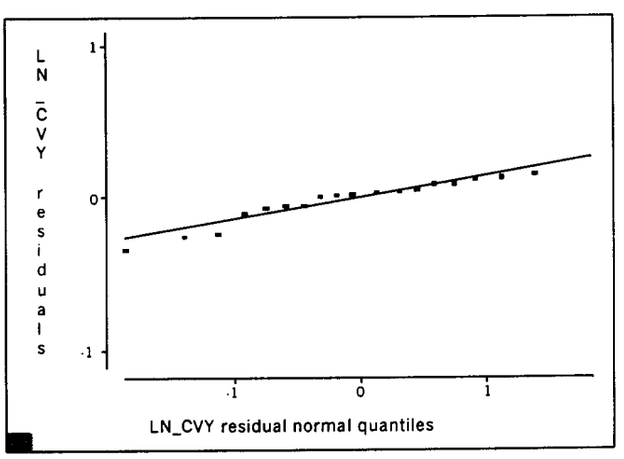
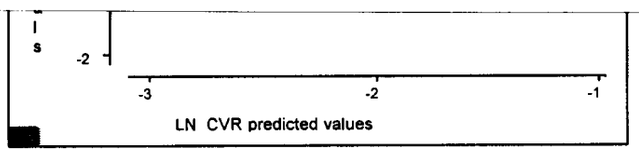
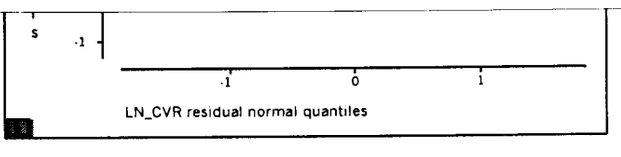


Fig 3:







9. Appendix - B

OHS 97 Workers data set (Official or strict definition of unemployment).

Results obtained from the SAS programs. (Please note: category '0' indicates that all the domain subclasses are included and not only a specific one.)

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	0	0	0	0	0	33105	7504	0.217176587	2417209	11130153	0.004373042	52227.60097	98341.3506	0.020135882	0.021606576	0.008835579
97	1	0	1	0	0	0	22261	4698	0.2004055	1624948	8108299	0.005363801	46333.71909	81963.68127	0.026764741	0.028513974	0.010108616
97	1	0	2	0	0	0	10844	2806	0.262177109	792261	3021854	0.006883268	23442.18107	44591.74787	0.026254267	0.029588968	0.014756422
97	1	0	0	0	0	1	17721	3282	0.18142816	1147325	6323853	0.004490088	29598.3946	61034.42713	0.02474857	0.025797741	0.009651462
97	1	0	0	0	0	2	15384	4222	0.264212313	1269884	4806300	0.005800762	31048.09815	52712.19143	0.021954926	0.024449562	0.010967312
97	1	0	1	0	0	1	11873	2037	0.166883809	765376	4586282	0.005405693	26010.26602	51596.42707	0.032391956	0.033983632	0.011250163
97	1	0	1	0	0	2	10388	2661	0.244056609	859572	3522017	0.007135486	27490.65273	44810.80432	0.029237013	0.03198181	0.012723052
97	1	0	2	0	0	1	5848	1245	0.219817677	381949	1737571	0.007596377	13956.20353	28252.21556	0.034557626	0.036539466	0.016259606
97	1	0	2	0	0	2	4996	1561	0.319487308	410312	1284283	0.009076794	13958.71439	22857.49295	0.028410501	0.034019749	0.017797864
97	1	0	0	0	1	0	22606	6408	0.281014557	2088753	7432900	0.005055938	49835.23776	93550.19262	0.017991728	0.023858844	0.012585961
97	1	0	0	0	2	0	5755	831	0.152525126	209235	1371804	0.008269506	14749.15768	42250.20162	0.054217338	0.070491023	0.030799011
97	1	0	0	0	3	0	1161	115	0.098898299	41944	424112	0.012920395	6055.392943	24256.1123	0.130643248	0.144368707	0.05719272
97	1	0	0	0	4	0	3583	150	0.04064346	77277	1901337	0.003999025	7823.874268	53841.7158	0.098392837	0.101244663	0.028317822
97	1	0	1	0	1	0	13307	3700	0.277989056	1317282	4738611	0.006559193	43723.17784	80832.52602	0.023595148	0.033191965	0.017058274
97	1	0	2	0	1	0	9299	2708	0.28633569	771471	2694289	0.007444732	23067.98487	42741.52595	0.026000014	0.02990129	0.015863747
97	1	0	1	0	2	0	4475	748	0.169691452	194602	1146798	0.009221961	14393.21774	40963.1595	0.054345465	0.073962409	0.035719598
97	1	0	2	0	2	0	1280	83	0.065032741	14633	225006	0.011140544	2540.453054	8880.726555	0.1713067	0.173614156	0.039468854
97	1	0	1	0	3	0	1142	113	0.098817545	41486	419826	0.013053026	6057.263049	24162.34899	0.132092195	0.146006786	0.057553245
97	1	0	2	0	3	0	19	2	0.106808567	458	4286	0	0	0	0	0	0
97	1	0	1	0	4	0	3337	137	0.03969786	71578	1803064	0.003922593	7194.555421	50736.54984	0.098811186	0.100513784	0.028139067
97	1	0	2	0	4	0	246	13	0.057992967	5699	98272	0.028783916	3225.749712	10306.59308	0.496334601	0.56600967	0.104877818
97	1	0	0	0	1	1	11777	2784	0.23763836	988873	4161252	0.00559734	27543.81085	53578.43986	0.023554024	0.02785374	0.012875559
97	1	0	0	0	1	2	10829	3624	0.33618529	1099880	3271649	0.006734204	29266.3455	47468.47969	0.020031227	0.026608666	0.014509038
97	1	0	0	0	2	1	3140	372	0.128479398	100382	781305	0.009379747	8665.35697	24412.52397	0.073005848	0.08632419	0.031245842
97	1	0	0	0	2	2	2615	459	0.184340655	108853	590499	0.010969143	8109.793774	18693.07568	0.059504739	0.07450226	0.0316564
97	1	0	0	0	3	1	739	64	0.086660971	23627	272639	0.013749685	3959.618085	15675.29754	0.158660636	0.167587329	0.057494622
97	1	0	0	0	3	2	422	51	0.120924588	18317	151473	0.01828524	3069.249763	9200.851404	0.151211925	0.16756518	0.060742712
97	1	0	0	0	4	1	2065	62	0.031067566	34443	1108657	0.004391151	4971.956388	31566.53086	0.141341981	0.14435196	0.028472753
97	1	0	0	0	4	2	1518	88	0.054036501	42834	792679	0.006564561	5252.046241	23935.8009	0.121483818	0.122615056	0.030196072
97	1	0	1	0	1	1	6873	1578	0.233818675	615730	2633364	0.007195452	23703.66142	45216.43301	0.030773642	0.038496859	0.017170595
97	1	0	1	0	1	2	6434	2122	0.333239922	701552	2105247	0.008927784	25557.36868	40567.70966	0.02679086	0.036429744	0.019269813
97	1	0	2	0	1	1	4904	1206	0.244221713	373143	1527887	0.008447066	13784.12318	26701.04725	0.034587693	0.036940568	0.017475797
97	1	0	2	0	1	2	4395	1502	0.341501402	398328	1166402	0.009520214	13648.6305	21581.98574	0.027877524	0.034264805	0.01850304
97	1	0	1	0	2	1	2385	339	0.145759442	93998	644884	0.010668491	8404.143585	23464.19755	0.073192454	0.089407795	0.036385163
97	1	0	1	0	2	2	2090	409	0.200440456	100604	501914	0.011996717	7808.60474	18032.36066	0.059851775	0.077617306	0.035927177

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	0	2	0	2	1	755	33	0.046793924	6384	136421	0.012315674	1698.514798	5516.018039	0.263189593	0.266071564	0.040433773
97	1	0	2	0	2	2	525	50	0.093120587	8249	88585	0.020970961	1891.411717	4693.868597	0.225202196	0.229287695	0.052987241
97	1	0	1	0	3	1	725	62	0.086016069	23169	269362	0.013915128	3953.644743	15585.9085	0.161773583	0.170640619	0.057862406
97	1	0	1	0	3	2	417	51	0.121734758	18317	150464	0.018351645	3065.892124	9195.46308	0.150751066	0.16738187	0.061113865
97	1	0	2	0	3	1	14	2	0.139657331	458	3278	0	0	0	0	0	0
97	1	0	2	0	3	2	5	0	0	0	1008	0	0	0	.	.	0
97	1	0	1	0	4	1	1890	58	0.031269917	32479	1038673	0.004559308	4813.459398	29938.13098	0.14580494	0.148201225	0.028823446
97	1	0	1	0	4	2	1447	79	0.051149944	39099	764392	0.006335423	4858.334515	22942.59409	0.123859822	0.124258579	0.030014192
97	1	0	2	0	4	1	175	4	0.028064383	1964	69985	0.016473633	1298.668653	7302.633134	0.586994295	0.661211248	0.104346252
97	1	0	2	0	4	2	71	9	0.132037034	3735	28288	0.05538339	1773.859472	1848.366028	0.419453458	0.474924863	0.06534155
97	1	1	0	0	0	0	5335	606	0.118159744	185061	1566189	0.008358207	13823.9698	30958.75891	0.070736503	0.074699715	0.019766931
97	1	2	0	0	0	0	2819	875	0.290738303	303402	1043555	0.018226476	20422.40952	35733.27628	0.062690315	0.067311495	0.03424186
97	1	3	0	0	0	0	1724	336	0.1854019	47209	254629	0.017273045	4160.916749	8820.988505	0.093165413	0.088138597	0.034642446
97	1	4	0	0	0	0	2821	624	0.203465737	156583	769578	0.012917579	9326.986879	23569.73768	0.063487735	0.059565863	0.030626834
97	1	5	0	0	0	0	5462	1361	0.228219463	474734	2080165	0.01113769	24779.86273	48127.83597	0.048802543	0.052197342	0.023136545
97	1	6	0	0	0	0	2798	686	0.240685374	190619	791984	0.01281856	11402.44297	21981.18526	0.053258573	0.059818013	0.027754594
97	1	7	0	0	0	0	6736	1553	0.216847661	670552	3092273	0.009544372	32243.91551	50205.74779	0.044014182	0.048085622	0.016235873
97	1	8	0	0	0	0	2961	801	0.244494269	178189	728807	0.013375113	10314.64103	23515.06747	0.05470522	0.057885884	0.032265133
97	1	9	0	0	0	0	2449	662	0.262600315	210861	802972	0.013835001	11185.71543	30484.92199	0.052684632	0.053047914	0.037965129
97	1	1	1	0	0	0	4354	570	0.12934128	178564	1380563	0.009288908	13598.01861	30021.37324	0.071817044	0.076152148	0.021745742
97	1	1	2	0	0	0	981	36	0.034998948	6497	185626	0.009279342	1737.075357	5902.536693	0.265132035	0.267377296	0.031797981
97	1	2	1	0	0	0	1646	430	0.244196718	162474	665339	0.025261837	18198.40217	26799.90877	0.103448717	0.112008331	0.040280069
97	1	2	2	0	0	0	1173	445	0.37261196	140928	378216	0.022951586	9131.949196	18960.59107	0.061596483	0.064798747	0.050131625
97	1	3	1	0	0	0	1255	293	0.233375875	41394	177372	0.021361072	3771.821774	6552.512963	0.091530763	0.091119417	0.036942274
97	1	3	2	0	0	0	469	43	0.075261279	5815	77258	0.017893163	1365.885931	4892.189831	0.237747258	0.234909393	0.063322907
97	1	4	1	0	0	0	2122	503	0.216331409	126035	582600	0.015299008	8498.946334	20544.51589	0.070720233	0.067433353	0.035263479
97	1	4	2	0	0	0	699	121	0.163377812	30548	186978	0.021572033	3863.92388	10277.34183	0.132037717	0.126486982	0.054965622
97	1	5	1	0	0	0	3203	655	0.190487101	260792	1369081	0.014165656	19903.56235	38119.09345	0.074365432	0.076319575	0.027842823
97	1	5	2	0	0	0	2259	706	0.300867277	213942	711084	0.01670471	14331.11527	23201.27639	0.055521856	0.066986038	0.032628051
97	1	6	1	0	0	0	1156	261	0.220529746	76852	348489	0.019665104	7735.281407	14694.00261	0.089172117	0.10065151	0.042164938
97	1	6	2	0	0	0	1642	425	0.256523225	113767	443495	0.016461676	8396.540032	15964.78471	0.064172265	0.073804855	0.035997665
97	1	7	1	0	0	0	6526	1535	0.221653691	660986	2982066	0.009754929	31930.86934	48502.78844	0.044009773	0.048307943	0.016264828
97	1	7	2	0	0	0	210	18	0.086802317	9566	110207	0.028029929	3231.758966	11678.15742	0.322916826	0.337830398	0.105965723
97	1	8	1	0	0	0	1421	359	0.219114968	86357	394115	0.017909888	7094.039489	15357.16493	0.081737402	0.082148225	0.038966166
97	1	8	2	0	0	0	1540	442	0.274379575	91833	334692	0.019390341	7260.621006	11988.43406	0.07066977	0.07906361	0.035819301
97	1	9	1	0	0	0	578	92	0.150926385	31494	208673	0.020437052	4286.567226	19395.16764	0.135410729	0.136105969	0.092945045
97	1	9	2	0	0	0	1871	570	0.301811929	179366	594298	0.015556624	10357.93133	19779.97117	0.051544101	0.057747375	0.033282909

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	1	0	0	0	1	2977	246	0.089061403	31094	910511	0.007924057	7394.002556	19987.90854	0.088972961	0.091181111	0.021963391
97	1	1	0	0	0	2	2358	360	0.158567292	103969	655678	0.012588807	8897.995603	18223.38414	0.079390947	0.085583031	0.027793172
97	1	2	0	0	0	1	1446	418	0.266409485	149448	560972	0.017837825	9801.553923	20053.09801	0.066956418	0.065584892	0.035747036
97	1	2	0	0	0	2	1373	457	0.31901902	153953	482583	0.023393363	13058.13054	19305.49071	0.073329054	0.08481885	0.040004495
97	1	3	0	0	0	1	965	146	0.140432216	21070	150037	0.017242567	2444.098125	6684.319648	0.122782131	0.115999103	0.044551256
97	1	3	0	0	0	2	759	190	0.249910106	26139	104593	0.0240455	2545.504296	4525.019331	0.096216596	0.097384062	0.043263167
97	1	4	0	0	0	1	1469	256	0.158674308	68829	433776	0.013646896	5608.971845	15257.17397	0.086005709	0.081491288	0.035172937
97	1	4	0	0	0	2	1352	368	0.261325552	87754	335802	0.016545078	5601.484968	11603.16989	0.063312134	0.063831927	0.034553601
97	1	5	0	0	0	1	2893	629	0.199970453	231908	1159713	0.011783989	14123.46511	28539.61654	0.058928648	0.060901076	0.02460921
97	1	5	0	0	0	2	2569	732	0.263811461	242826	920452	0.013999545	14598.15083	26221.2353	0.053066478	0.06011777	0.028487335
97	1	6	0	0	0	1	1512	309	0.202661245	93449	461110	0.014368867	7057.29234	14676.07914	0.070900911	0.075520107	0.031827699
97	1	6	0	0	0	2	1286	377	0.293676388	97170	330873	0.016114338	6497.290564	10758.62615	0.054871072	0.066865395	0.03251584
97	1	7	0	0	0	1	3648	685	0.183214053	327414	1787055	0.009663869	18717.85981	32073.12838	0.052746332	0.057168842	0.017947473
97	1	7	0	0	0	2	3088	868	0.262897553	343138	1305217	0.012777804	18312.04324	27834.80614	0.048603739	0.05336634	0.021325801
97	1	8	0	0	0	1	1612	306	0.17222659	74944	435150	0.013006747	5572.167468	15649.23004	0.075521132	0.074350703	0.03596285
97	1	8	0	0	0	2	1349	495	0.351582538	103245	293657	0.018658533	6646.64507	11000.33368	0.053070134	0.064377499	0.037459741
97	1	9	0	0	0	1	1199	287	0.233052807	99171	425529	0.01777784	7563.337945	18499.0742	0.076282452	0.07626585	0.043473133
97	1	9	0	0	0	2	1250	375	0.29591217	111690	377443	0.016348578	6450.539788	14818.47098	0.055248076	0.057754009	0.039260181
97	1	1	0	0	1	0	959	220	0.225676101	75771	335752	0.024651818	9930.482818	20369.08247	0.109235397	0.131058641	0.060666974
97	1	1	0	0	2	0	3621	363	0.111867795	98266	878413	0.007860167	8820.170152	31230.10763	0.070263	0.089758028	0.035552889
97	1	1	0	0	3	0	65	5	0.092238429	2010	21790	0.046911711	927.7890788	6328.640994	0.50859183	0.461622287	0.290442301
97	1	1	0	0	4	0	690	18	0.027293668	9013	330235	0.007227353	2440.935659	23234.55839	0.264799611	0.270814383	0.070357714
97	1	2	0	0	1	0	2118	757	0.351420935	268559	764208	0.019869647	19582.55552	31566.07491	0.056540875	0.072917237	0.041305611
97	1	2	0	0	2	0	498	115	0.227178477	33196	146125	0.034697966	6206.192557	13409.88287	0.152734388	0.186953533	0.091769835
97	1	2	0	0	3	0	18	1	0.05848942	399	6817	0.056588887	398.7216511	2733.632947	0.967506386	1	0.401003065
97	1	2	0	0	4	0	185	2	0.009870534	1248	126406	0.006567453	849.8383834	14919.60244	0.665359434	0.681129476	0.118029685
97	1	3	0	0	1	0	474	114	0.238387907	20500	85993	0.031421092	2938.616388	7051.66363	0.131806569	0.14334872	0.082002432
97	1	3	0	0	2	0	1041	215	0.202030418	25196	124716	0.019038248	2841.886925	7332.822359	0.094234563	0.112789605	0.058796328
97	1	3	0	0	3	0	2	0	0	0	204	0	0	0	.	.	0
97	1	3	0	0	4	0	207	7	0.034601889	1513	43716	0.015114224	672.0554894	5643.023523	0.436803441	0.444286234	0.129083101
97	1	4	0	0	1	0	2489	598	0.237179056	146514	617734	0.013842554	9360.239564	20275.32264	0.058363307	0.063886526	0.032822104
97	1	4	0	0	2	0	80	9	0.107825879	2752	25518	0.02797965	1063.446916	4836.919515	0.259489187	0.38649059	0.189546154
97	1	4	0	0	3	0	11	1	0.092669769	101	1088	0.044356136	100.8138101	568.4560162	0.478647314	1	0.522534438
97	1	4	0	0	4	0	241	16	0.057625341	7217	125238	0.01643173	2068.422948	11570.12187	0.285147633	0.286609157	0.092385148
97	1	5	0	0	1	0	4156	1229	0.286625042	417429	1456358	0.011984963	22993.49053	40197.19857	0.04181408	0.055083627	0.027601173
97	1	5	0	0	2	0	94	23	0.271666849	13184	48529	0.093473624	6491.887952	11164.46034	0.344074459	0.49241937	0.230058611
97	1	5	0	0	3	0	852	92	0.104109935	32999	316960	0.016375782	5678.562328	20061.42228	0.157293174	0.172084243	0.063293134

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	5	0	0	4	0	360	17	0.043059509	11123	258318	0.013002948	3573.427112	29280.16722	0.301976216	0.321263795	0.113349466
97	1	6	0	0	1	0	2589	664	0.257375583	182740	710012	0.012944523	10876.15016	20529.34608	0.050294295	0.059517124	0.028914068
97	1	6	0	0	2	0	38	9	0.229834164	2300	10008	0.074964207	796.4847416	3058.740415	0.326166508	0.346261963	0.305621943
97	1	6	0	0	3	0	6	0	0	0	1971	0	0	0	.	.	0
97	1	6	0	0	4	0	165	13	0.079706386	5579	69992	0.033489734	2467.58018	9008.717665	0.420163745	0.442313866	0.128710914
97	1	7	0	0	1	0	4775	1392	0.286276725	599997	2095862	0.010173082	32087.45772	61993.54302	0.035535835	0.053479404	0.029579018
97	1	7	0	0	2	0	343	89	0.250641759	32643	130239	0.031725945	6943.985574	20316.29048	0.126578845	0.212722359	0.155991945
97	1	7	0	0	3	0	164	12	0.083099626	5768	69411	0.019173921	1838.612823	11530.19669	0.230734144	0.318758983	0.166114649
97	1	7	0	0	4	0	1454	60	0.040343522	32144	796760	0.006252208	5091.322853	29714.65211	0.154974272	0.158390526	0.037294348
97	1	8	0	0	1	0	2650	777	0.279913571	170330	608511	0.012952867	10170.38362	20165.05586	0.046274524	0.059709745	0.03313838
97	1	8	0	0	2	0	38	8	0.220178859	1697	7706	0.073127165	1010.07482	2449.106311	0.33212619	0.595281097	0.317798469
97	1	8	0	0	3	0	41	3	0.068716525	356	5174	0.047339957	106.2144523	2259.490173	0.68891664	0.298716781	0.436664264
97	1	8	0	0	4	0	232	13	0.054056045	5806	107416	0.016918461	1702.73161	10345.04921	0.312979997	0.293246972	0.096308399
97	1	9	0	0	1	0	2396	657	0.272805203	206914	758470	0.013637743	11188.27542	24339.46564	0.049990772	0.054071989	0.032090234
97	1	9	0	0	2	0	2	0	0	0	549	0	0	136.2709791	.	.	0.248081968
97	1	9	0	0	3	0	2	1	0.448669426	312	696	0	0	0	0	0	0
97	1	9	0	0	4	0	49	4	0.084008354	3634	43257	0.025052973	1457.184861	6504.913924	0.298220025	0.400992977	0.150378637
97	1	0	1	0	0	0	21858	4594	0.199417113	1592953	7988046	0.005400697	45983.20654	79682.69808	0.027082414	0.028866642	0.009975243
97	1	0	2	0	0	0	11247	2910	0.26232575	824256	3142107	0.006869313	24235.68978	46650.14562	0.026186193	0.029403127	0.014846773
97	1	0	1	0	0	1	11659	2000	0.166295423	751430	4518646	0.005454325	25822.81985	50307.16093	0.032799008	0.034364896	0.011133238
97	1	0	1	0	0	2	10199	2594	0.242555754	841523	3469400	0.007180894	27306.38094	43907.73166	0.029605128	0.032448767	0.012655713
97	1	0	2	0	0	1	6062	1282	0.219307186	395895	1805207	0.007492242	14366.61883	29594.36924	0.034163232	0.036288975	0.016393892
97	1	0	2	0	0	2	5185	1628	0.32041343	428361	1336900	0.009073686	14300.69714	23506.0159	0.028318683	0.033384711	0.017582481
97	1	0	1	0	1	0	12918	3602	0.278362521	1287902	4626708	0.006600139	43324.46393	79521.1732	0.023710587	0.033639565	0.017187421
97	1	0	2	0	1	0	9688	2806	0.285387098	800851	2806193	0.007329576	23796.28445	44025.40337	0.025682928	0.02971374	0.01568866
97	1	0	1	0	2	0	4461	745	0.169270474	193723	1144458	0.00925209	14422.78928	41054.11821	0.05465862	0.074450572	0.035872093
97	1	0	2	0	2	0	1294	86	0.068228935	15512	227345	0.011289582	2584.77955	8915.991404	0.165466187	0.166635936	0.039217815
97	1	0	1	0	3	0	1134	111	0.098150527	41073	418471	0.013096322	6056.424045	24155.68044	0.133430987	0.14745463	0.057723685
97	1	0	2	0	3	0	27	4	0.154370686	871	5641	0	0	0	0	0	0
97	1	0	1	0	4	0	3345	136	0.039065032	70255	1798409	0.00394347	7209.457183	50629.54053	0.100946289	0.102618554	0.028152405
97	1	0	2	0	4	0	238	14	0.068222628	7022	102928	0.026569362	3205.804243	9895.440383	0.389450875	0.456537397	0.096139737
97	1	0	1	0	1	1	6655	1540	0.234847262	602539	2565663	0.007289428	23488.41553	44267.65326	0.031039016	0.038982404	0.017253885
97	1	0	1	0	1	2	6263	2062	0.332531885	685363	2061045	0.00898327	25352.13229	39919.14046	0.027014763	0.036990803	0.0193684
97	1	0	2	0	1	1	5122	1244	0.242126367	386334	1595589	0.008262656	14190.81946	28073.68712	0.034125387	0.03673199	0.017594565
97	1	0	2	0	1	2	4566	1562	0.342405185	414517	1210604	0.009406564	13932.485	21984.90576	0.027472025	0.033611362	0.018160277
97	1	0	1	0	2	1	2378	340	0.146021004	94027	643925	0.01068443	8415.141115	23525.70669	0.073170499	0.089497506	0.036534867
97	1	0	1	0	2	2	2083	405	0.199180375	99696	500534	0.012039366	7825.374792	18086.48772	0.060444541	0.078491994	0.036134413

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	0	2	0	2	1	762	32	0.046258699	6355	137380	0.012428826	1725.056599	5624.876033	0.268680842	0.271447969	0.040943938
97	1	0	2	0	2	2	532	54	0.10177813	9157	89965	0.020836874	1910.590755	4700.284304	0.204728407	0.208659053	0.052245411
97	1	0	1	0	3	1	721	62	0.086248416	23169	268636	0.013952779	3953.644743	15583.90169	0.161774323	0.170640619	0.058011234
97	1	0	1	0	3	2	413	49	0.119489572	17904	149835	0.018440891	3065.892124	9195.46308	0.154330546	0.171243374	0.061370621
97	1	0	2	0	3	1	18	2	0.114343905	458	4003	0	0	0	0	0	0
97	1	0	2	0	3	2	9	2	0.252225419	413	1638	0	0	0	0	0	0
97	1	0	1	0	4	1	1905	58	0.030463861	31695	1040422	0.004453482	4691.667803	29915.0171	0.146189023	0.148024151	0.028752759
97	1	0	1	0	4	2	1440	78	0.050871114	38560	757987	0.006367223	4848.591445	23040.46549	0.125163811	0.125742709	0.03039693
97	1	0	2	0	4	1	160	4	0.04027265	2748	68235	0.01515593	1248.302813	7014.547475	0.37633306	0.454258038	0.102799864
97	1	0	2	0	4	2	78	10	0.123195766	4274	34693	0.045537848	1773.859472	1829.156363	0.36963809	0.41503563	0.052724538
97	1	1	1	0	0	0	4342	569	0.129389461	178405	1378825	0.009298994	13596.59465	30016.89232	0.071868248	0.076211761	0.021769902
97	1	1	2	0	0	0	993	37	0.035519485	6655	187364	0.009226983	1741.145331	5880.630415	0.259772421	0.261626644	0.03138611
97	1	2	1	0	0	0	1650	440	0.247678643	165976	670128	0.025150057	18265.4855	26706.59404	0.1015431	0.110048748	0.039852994
97	1	2	2	0	0	0	1169	435	0.368010212	137425	373428	0.023101763	8835.627343	18136.69089	0.062774787	0.064294083	0.048568147
97	1	3	1	0	0	0	1255	285	0.220712305	38836	175956	0.019377607	3326.020566	6510.64268	0.08779577	0.08564371	0.037001638
97	1	3	2	0	0	0	469	51	0.106429645	8373	78674	0.034491096	2639.325319	4916.409125	0.32407414	0.315209459	0.062490924
97	1	4	1	0	0	0	2000	477	0.217375785	119520	549831	0.015880227	8367.817278	19558.39785	0.073054258	0.070011837	0.035571627
97	1	4	2	0	0	0	821	147	0.168661182	37063	219747	0.01967461	4091.005736	11609.56038	0.116651679	0.11038062	0.052831587
97	1	5	1	0	0	0	3017	596	0.185191305	242888	1311549	0.01426492	19248.38017	36809.59814	0.077028023	0.079248126	0.028065739
97	1	5	2	0	0	0	2445	765	0.301641755	231847	768616	0.016696581	15324.51624	24307.63789	0.055352354	0.066097632	0.031625204
97	1	6	1	0	0	0	1126	263	0.227735532	77921	342154	0.020345	7903.279466	13829.60359	0.08933608	0.101427342	0.040419248
97	1	6	2	0	0	0	1672	423	0.250535409	112698	449830	0.015499274	7983.986838	16619.84125	0.061864605	0.070843908	0.036946958
97	1	7	1	0	0	0	6556	1535	0.220498248	659668	2991717	0.009767501	31956.81985	48766.88622	0.044297406	0.048443761	0.016300633
97	1	7	2	0	0	0	180	18	0.108235645	10884	100555	0.035248883	4148.122593	8871.927472	0.325667974	0.381132382	0.088229244
97	1	8	1	0	0	0	1407	360	0.221522149	86679	391290	0.01808303	6944.152266	14268.93618	0.081630799	0.080113075	0.036466402
97	1	8	2	0	0	0	1554	441	0.271126254	91510	337517	0.019351204	7319.09565	12284.52591	0.071373405	0.079981513	0.036396713
97	1	9	1	0	0	0	505	69	0.130579552	23060	176596	0.01856503	3529.250195	13495.95519	0.142174096	0.153047774	0.076422875
97	1	9	2	0	0	0	1944	593	0.299821265	187801	626376	0.015805663	10639.8059	23602.13312	0.052716953	0.056654738	0.037680462

OHS: October Households Survey

STR: Official strict definition of unemployment

U / R: Urban / Rural

Type: Urban formal, Urban informal, Tribal, Commercial farms, Other non-urban

N: Population number in subclass

n: Sample size of subclass

R: Estimated ratio

MSWY: Estmated number of unemploued

MSWX: Estimated number of economic active

SE-R: estimated standard error of R

SE-WY: Estimated standard error of MSWY

SE-WX: Estimated standard error of MSWX

CV-R: Estimated coefficient of relative variation of R

CV-WY: Estimated coefficient of relative variation of MSWY

CV-WY: Estimated coefficient of relative variation of MSWX