International
Labour
Organization

ILO-IPEC Interactive Sampling Tools No. 7

Calculation of sampling errors

Version 1

December 2014

**International
Programme on
the Elimination
of Child Labour
(IPEC)**

**Fundamental Principles and Rights at Work (FPRW) Branch
Governance and Tripartism Department**

ILO-IPEC

ILO-IPEC Interactive Sampling Tools No. 7 - Calculation of sampling errors / International Labour Office, International Programme on the Elimination of Child Labour (IPEC) - Geneva: ILO, 2014

ACKNOWLEDGEMENTS

This publication was elabourated by Mr. Farhad Mehran, consultant, for ILO-IPEC and coordinated by Mr. Federico Blanco Allais from IPEC Geneva Office.

Funding for this ILO publication was provided by the United States Department of Labour (Projects GLO/13/21/USA & GLO/10/55/USA).

This publication does not necessarily reflect the views or policies of the United States Department of Labour, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government.

The designations employed in ILO publications, which are in conformity with United Nations practice, and the presentation of material therein do not imply the expression of any opinion whatsoever on the part of the International Labour Office concerning the legal status of any country, area or territory or of its authorities, or concerning the delimitation of its frontiers.
The responsibility for opinions expressed in signed articles, studies and other contributions rests solely with their authors, and publication does not constitute an endorsement by the International Labour Office of the opinions expressed in them.
Reference to names of firms and commercial products and processes does not imply their endorsement by the International Labour Office, and any failure to mention a particular firm, commercial product or process is not a sign of disapproval.
ILO publications and electronic products can be obtained through major booksellers or ILO local offices in many countries, or direct from ILO Publications, International Labour Office, CH-1211 Geneva 22, Switzerland. Catalogues or lists of new publications are available free of charge from the above address, or by email: pubvente@ilo.org or visit our website: www.ilo.org/publns.

Visit our website: www.ilo.org/ipec

Available in electronic PDF format only.

Photocomposed by ILO-IPEC Geneva.

1. Introduction

This document describes the use of the template “Sampling Errors” of the SIMPOC Interactive Sampling Tools. The template assists the user to calculate the sampling errors of the main survey estimates as well as approximate values for any other estimates. The calculations of the sampling errors are based on the sampling variations of replicates constructed using data at the PSU level.

The template is divided into three parts: Input values, Output values and Intermediary calculations. The present document describes the contents and use of each part. Initially, however, the general methodology is described in Section 2 before turning to its application in the template with Input values in Section 3, Output values in Section 4, and Intermediary calculations in Section 5.

2. Methodology

Like in all sample surveys, the results of child labour surveys are subject to sampling errors. Sampling errors arise due to the fact that the survey does not cover all elements of the population, but only a selected portion. The sampling error of an estimate is based on the difference between the estimate and the value that would have been obtained on the basis of a complete count of the population under otherwise identical conditions.

Information on sampling errors is used for interpreting the survey results. It provides an assessment of the precision of the estimates and on the degree of confidence that may be attached to them. In the same vein, it allows decision on the degree of detail with which the survey data may be meaningfully tabulated and analysed. Information on sampling errors is also used for determining whether the survey estimates of change over time or the estimates of differences between two or more population subgroups are statistically significant. Finally, information on sampling errors may be used for future sample design. Rational decisions on the choice of sample size, sample allocation among strata, clustering and estimation procedures, can only be made on the basis of detail knowledge of their effect on the magnitude of sampling errors in the resulting statistics obtained from the survey.

The calculation of the sampling variance of survey estimates for complex multi-stage designs is generally based on the following principle: the variance contributed by the later stages of sampling is, under broad conditions, reflected in the observed variation among the sample results for first-stage units. Thus, the sampling variance of a variety of statistics, such as totals, means, ratios, proportions, and their differences can be obtained on the basis of totals calculated for primary sampling units (PSUs).[1]

[1] Verma, Vijay, *Sampling Methods*, Manual for Statistical Trainers Number 2, Statistical Institute for Asia and the Pacific (SIAP), Tokyo, Revised 2002.

Suppose that the results of the survey give an estimated total number of children (x) and working children (y). Let m$_h$ be the number of sample PSUs in stratum h selected from a total of M$_h$ sample PSUs in stratum h. The survey estimates of the number of working children and of the total number of children may be expressed respectively as

$$y=\sum_h\sum_i y_h$$

$$x=\sum_h\sum_i x_h$$

where y$_{hi}$ and x$_{hi}$ are the corresponding sum of sample results for sample PSUi, The sampling variance of the estimates y and x can be calculated by

$$var(y)=\sum_h(1-f_h)\frac{m_h}{m_h-1}\sum_i(y_h-\frac{y_h}{m_h})^2$$

$$var(x)=\sum_h(1-f_h)\frac{m_h}{m_h-1}\sum_i(x_h-\frac{x_h}{m_h})^2$$

where f$_h$=m$_h$/M$_h$ is the sampling fraction and $y_h=\sum_i y_{hi}$ and $x_h=\sum_i x_{hi}$.

The sample principle applies for the calculation of the sampling variance of means, proportions, percentages, and ratios where both the numerator and denominator are sample estimates such as for the calculation of the sampling variance of the percentage of working children,

$$r=\frac{y}{x}$$

The sampling variance is derived by Taylor linearization of the statistic,

$$var(z)=\sum_h(1-f_h)\frac{m_h}{m_h-1}\sum_i(z_h-\frac{z_h}{m_h})^2$$

where $z_{hi}=\frac{1}{x}(y_{hi}-rx_{hi})$.

Taylor linearization can be applied for the calculation of the sampling variance of more complex statistics such as differences of ratios, ratio of ratios, regression coefficients, etc.[2]

[2] It should be mentioned that there are other methods of variance estimation for complex designs. Some of these alternative methods are based on comparison among replications of the full sample, such as jack-knife repeated replications, balanced repeated replications and bootstrapping. A major feature of these procedures is that, under general conditions for their application, the same and relatively simple variance estimation formula holds for statistics of any complexity.

3. Input values

The input data for calculating sampling errors are the weighted values by PSU of the variables for which the sample errors are to be calculated. If the target variables are in the form of ratios, the weighted values sample of the numerator and denominator should be given separately.

Table 1 shows a numerical example of input values to be entered in the template. Each row represents a sample PSU. Column (1) identifies the PSU with its PSU code number. Column (2) specifies the stratum in which it is located. Column (3) gives the sampling weight of the sample households and individuals assumed to be constant in the PSU. The next columns (4) to (10) are for entering the sample values of the variables for which sampling errors should be calculated. If the sampling weights of the households and individuals in the PSU are not constant, the values in column (3) should be set to 1 and the values entered in columns (4) to (10) should be weighted sum of the variables in question.

Table 1. Input values: Numerical illustration

INPUT VALUES									
								rounding	3
Identifiers			Variables						
PSU code	Stratum code	Sampling weights	Number of children 5-17 yrs	Number of working children	Child labour	In hazardous work	Agr CL	Ind CL	Ser CL
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1111	111	660	11	0	0	0	0	0	0
1112	111	1054	13	0	0	0	0	0	0
1113	111	781	17	5	3	2	3	0	0
1114	111	754	10	3	3	1	2	1	0
1115	111	749	14	4	3	2	3	0	0
1116	111	826	15	2	2	0	1	0	1
1117	111	589	14	0	0	0	0	0	0
2111	211	1257	11	1	1	0	0	0	1
2112	211	2401	13	5	3	1	3	0	0
2113	211	2665	13	4	3	2	1	0	2
2114	211	2311	17	6	4	3	4	0	0

In the numerical example of Table 1, the variables for which sampling errors are to be calculated include total number of children 5-17 years old, the number of working children, the number children engaged in child labour and its breakdown by major branch of economic activity (agriculture, industry and services) and the number of children engaged hazardous work.

Row 1 refers to the input values for the PSU with code number 1111 in stratum 111. The sample households and individuals in this PSU have a constant sampling weight equal to 660. In total there were 11 children 5 to 17 years old in the sample households in this PSU, and none were working.

Similarly, row 2 gives the input values for the PSU with code number 1112 in the same stratum 111. The sample households and individuals in this PSU have a constant sampling weight equal to 1054. In total there were 13 children 5 to 17 years old in the sample households in this PSU, also none working.

In row 3 corresponding to the PSU with code number 1112 in the stratum 111 and with sampling weight 781, there were in total 17 children 5 to 17 years old in the sample households, 5 working, 3 engaged in child labour, all of them in agriculture. One child was working in hazardous conditions. And so on for the other rows.

As an option, one may specify the rounding rule to be used for the output values. This is given in the top right corner of the input values. In the numerical example in Table 1, the rounding rule is set to 3, meaning that in the output values, the sampling errors of levels should be rounded to 1000.

4. Output values

There are three sets of variables: sampling errors of the estimates of the variables of specified in the input values, sample errors of certain ratios of these variables and approximate sampling errors for general variables. These are in described in turn below.

Sampling errors of estimated levels

The sampling errors of the variables of interest specified in the input values are calculated and the results reported in the output values as shown in Table 2 below. Column (15) specifies the indicator. Column (16) gives the estimated value. Columns (17) and (18) give the standard error or standard deviation of the estimate and the corresponding relative standard error or standard deviation. Finally, columns (19) and (20) give the lower and upper limits of the confidence interval of the estimate at the 95% level of confidence

One use of the standard deviation is to assess the level of precision of the estimate. A low relative standard deviation indicates a high precision of the estimate. In general, the lower the relative standard deviation of an estimate, the higher is the precision of the estimate. The relative standard deviation of an estimate is the ratio of the standard deviation to the size of the estimate.

Table 2. ***Output values: Sampling errors of estimated levels***

OUTPUT VALUES					
				Confidence interval	
Indicator	Estimate	Standard error	Relative standard error	Lower	Upper
(15)	(16)	(17)	(18)	(19)	(20)
Number of children 5-17 years	2,919,600	119,100	4.1%	2,681,400	3,157,800
Number of working children	588,500	41,100	7.0%	506,300	670,700
Child labour	392,700	26,500	6.7%	339,700	445,700
- Agriculture	235,000	21,000	8.9%	193,000	277,000
- Industry	10,800	4,000	37.0%	2,800	18,800
- Services	146,800	18,900	12.9%	109,000	184,600
Hazardous work by children	114,700	16,400	14.3%	81,900	147,500

Thus, in this numerical example, the estimate of the total number of children 5 to 17 years old is 2,919,600 with standard error 119,100. The relative standard error of the estimate is 4.1%. The estimates of the number of working children and the number in child labour, with relative standard errors of 7.0% and 6.7%, respectively, are estimated less precisely than the total number of children 5-17 years old. The results also show that child labour in industry is estimated with the least precision (37.0%). Child labour in agriculture and in services and hazardous work by children are estimated with mid-level precision with relative standard errors around 8.9%, 12.9% and 14.3%, respectively.

Another use of the standard errors is for the calculation of confidence intervals. Under certain broad assumptions, it can be stated that the true value of the variable of interest lies in between the survey estimate and a multiple of the standard error, with certain degree of probability. Thus, referring to the results shown in Table 2, it can be stated, for example, that the true value of the total number of children 5 to 17 years old is within the interval,

$$2{,}919{,}600 - 2 \times 119{,}100 \leq \theta \leq 2{,}919{,}600 + 2 \times 119{,}100$$

$$2{,}681{,}400 \leq \theta \leq 3{,}157{,}800,$$

where the multiplicative factor 2 is the rounded value of the standard normal distribution corresponding to 95% confidence probability and 119,100 is the standard error of the survey estimate of the total number of children 5-17 years old, 2,919,600.

Sampling errors of estimated ratios

The next set of output values shown in Table 3 below gives the standard errors of estimated ratios. It shows that the percentage of working children and the child labour rate are estimated with standard errors of about 0.6 and 0.4 percentage points respectively. The estimates of other child labour indicators, in particular, the percentage of child labour in agriculture and services and the percentage of child labour in hazardous work have standard errors of about 0.2 percentage points. The standard error of the estimated percentage of child labour in industry is below 0.1 percentage point and reported here as 0.0%.

Table 3. Output values: Sampling errors of estimated ratios

OUTPUT VALUES					
				Confidence interval	
Indicator	Estimate	Standard error	Relative standard error	Lower	Upper
(15)	(16)	(17)	(18)	(19)	(20)
% working children	20.2%	0.6%		18.9%	21.4%
% child labour	13.5%	0.4%		12.6%	14.3%
% child labour in agriculture	59.8%	1.7%		56.5%	63.2%
% child labour in industry	2.8%	0.3%		2.2%	3.3%
% child labour in services	37.4%	1.6%		34.1%	40.6%
% child labour in hazardous work	29.2%	1.6%		26.0%	32.4%

The right columns of Table 3 give the confidence intervals of the estimated percentages. Thus, from the second row of the table, one notes that the estimated child labour rate lies at 95% confidence within the following interval,

$$13.5\% - 2 \times 0.4\% \leq \theta \leq 13.5\% + 2 \times 0.4\%$$

$$12.6\% \leq \theta \leq 14.3\%$$

Similar inferences may be made on the estimates of the other child labour indicators reported in Table 3.

Generalized sampling errors

As it is not practical to compute and report sampling variances and standard errors for every published statistics of a child labour survey, it is customary to give general variance estimates using the approximate relationship between the variance of an estimate and the level of the estimate, expressed by

$$\frac{\mathrm{var}(y)}{y^2} = a + b \times \frac{1}{y}$$

where the parameters a and b are estimated by linear regression.

The output values are given in Table 4 below. They are calculated on the basis of the regression fitted to the seven main results obtained earlier in Table 2. Thus, an estimated value of about 2,500,000 has an approximate standard error of 156,000 corresponding to a relative standard error of about 6.2%. Similarly, an estimated value of about 500,000 has an approximate standard error of 39,000 corresponding to a relative standard error of about 7.8%. For small estimates of around 50'000, the approximate standard error is about 9,000 corresponding a high relative standard error of about 18.0%.

Table 4. Output values: Generalized sampling errors

OUTPUT VALUES					
				Confidence interval	
Indicator	Estimate	Standard error	Relative standard error	Lower	Upper
(15)	(16)	(17)	(18)	(19)	(20)
Levels	Approximate standard errors				
	2,500,000	157,200	6.3%	2,185,600	2,814,400
	1,000,000	69,500	7.0%	861,000	1,139,000
	750,000	54,700	7.3%	640,600	859,400
	500,000	39,600	7.9%	420,800	579,200
	250,000	23,900	9.6%	202,200	297,800
	100,000	13,400	13.4%	73,200	126,800
	75,000	11,300	15.1%	52,400	97,600
	50,000	9,000	18.0%	32,000	68,000
	25,000	6,200	24.8%	12,600	37,400

The results may also be used as follows. Suppose, for example, the survey results indicate that the estimated number of girls in child labour is 190,000. The approximate standard error of the estimate may be calculated from Table 4 by interpolation,

$$Stderror = 13{,}400 + \frac{(23{,}900 - 13{,}400)}{(250{,}000 - 100{,}000)} \times (190{,}000 - 100{,}000)$$

$$= 19{,}600$$

4. Intermediary calculations

In line with the three sets of output values, there are three sets of intermediary calculations, one for calculating the sampling errors of estimated levels, the other for calculating the sampling errors of estimated ratios, and finally one for estimating the generalized variances.

In the next page, the intermediary calculations for deriving the sampling errors of the estimates of levels are presented. They involve five steps:

1. calculation of the weighted sums in each PSU (y_j) for the different target variables;
2. calculation of the corresponding weighted squared values (y_j^2);
3. aggregation of the weighted values over all sample PSUs in the same stratum to obtain y_h;
4. aggregation of the weighted squared values over all sample PSUs in the same stratum to obtain y^2_h; and
5. finally calculation of the standard errors within each stratum (σ_h)

INTERMEDIARY CALCULATIONS						
yj						
Number of children 5-17 yrs	Number of working children	Child labour	In hazardous work	Agr CL	Ind CL	Ser CL
(22)	(23)	(24)	(25)	(26)	(27)	(27)
7263	0	0	0	0	0	0
13704	0	0	0	0	0	0
13272	3903	2342	1561	2342	0	0
yj2						
Number of children 5-17 yrs	Number of working children	Child labour	In hazardous work	Agr CL	Ind CL	Ser CL
(28)	(29)	(30)	(31)	(32)	(32)	(33)
52749304	0	0	0	0	0	0
187803728	0	0	0	0	0	0
176133367	15236450	5485122	2437832	5485122	0	0

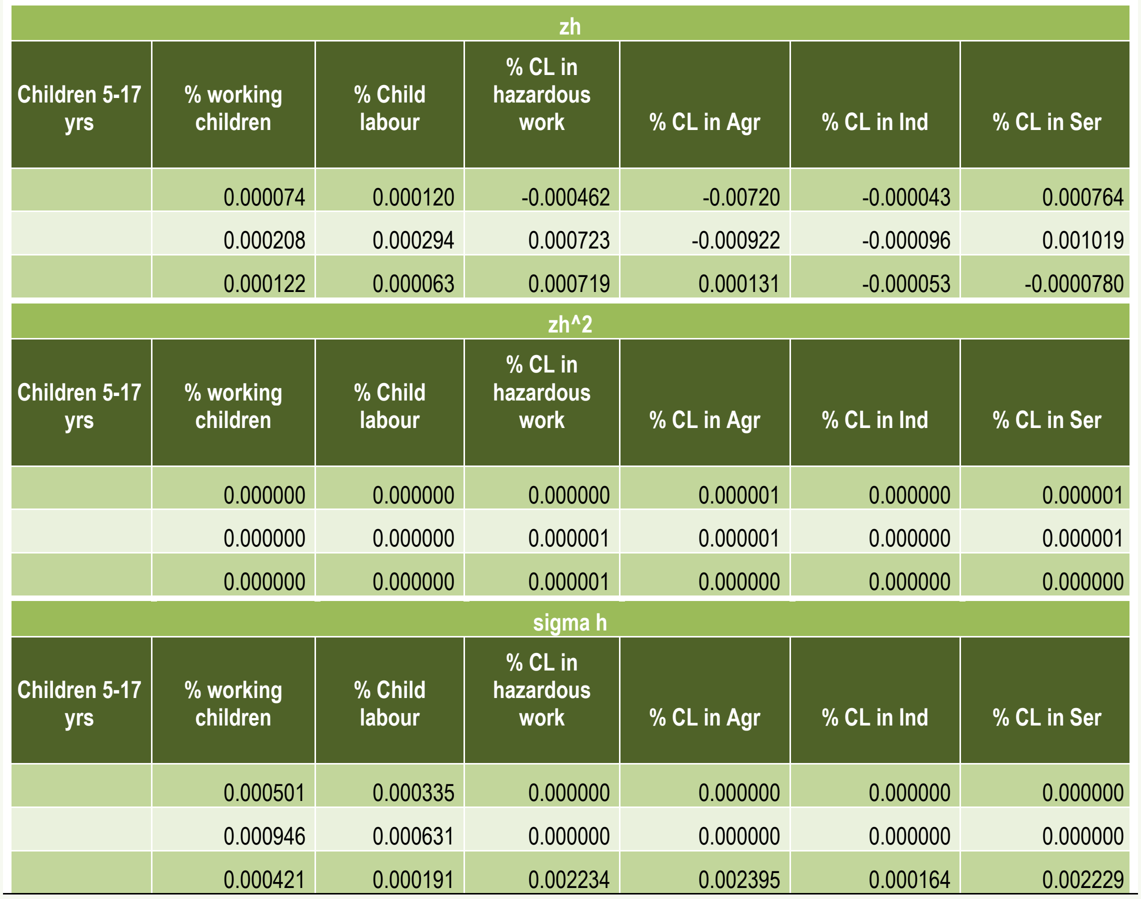
yh						
Number of children 5-17 yrs	Number of working children	Child labour	In hazardous work	Agr CL	Ind CL	Ser CL
72896	10815	8504	3814	6924	754	826
14375	2226	1305	128	600	0	705
247698	53132	26015	6824	16239	542	9234

yh^2						
Number of children 5-17 yrs	Number of working children	Child labour	In hazardous work	Agr CL	Ind CL	Ser CL
804916613	32067632	18385226	5252918	13496152	568756	681765
63373410	1327770	442430	16369	194212	0	248218
4572186356	299912139	58321151	6849633	36334921	293855	13807915

sigma h						
Number of children 5-17 yrs	Number of working children	Child labour	In hazardous work	Agr CL	Ind CL	Ser CL
7309	4233	3065	1924	2785	754	826
3952	345	149	128	373	0	407
22723	10940	3761	2003	4483	542	2950

zj						
Number of children 5-17 yrs	% working children	% Child labour	% CL in hazardous work	% CL in Agr	% CL in Ind	% CL in Ser
(28)	(23)	(24)	(25)	(26)	(27)	(27)
	-0.000501	-0.000335	0.000000	0.000000	0.000000	0.000000
	-0.000946	-0.000631	0.000000	0.000000	0.000000	0.000000
	0.000421	0.000191	0.002234	0.002395	-0.000164	-0.002229

zj2						
Children 5-17 yrs	% working children	% Child labour	% CL in hazardous work	% CL in Agr	% CL in Ind	% CL in Ser
(28)	(29)	(30)	(31)	(32)	(32)	(33)
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000001	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000005	0.000006	0.000000	0.000005



zh						
Children 5-17 yrs	% working children	% Child labour	% CL in hazardous work	% CL in Agr	% CL in Ind	% CL in Ser
	0.000074	0.000120	-0.000462	-0.00720	-0.000043	0.000764
	0.000208	0.000294	0.000723	-0.000922	-0.000096	0.001019
	0.000122	0.000063	0.000719	0.000131	-0.000053	-0.0000780

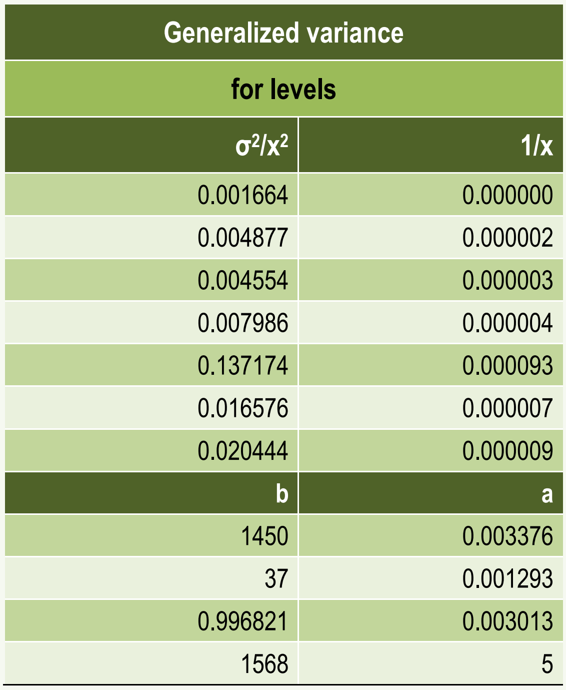
zh^2						
Children 5-17 yrs	% working children	% Child labour	% CL in hazardous work	% CL in Agr	% CL in Ind	% CL in Ser
	0.000000	0.000000	0.000000	0.000001	0.000000	0.000001
	0.000000	0.000000	0.000001	0.000001	0.000000	0.000001
	0.000000	0.000000	0.000001	0.000000	0.000000	0.000000

sigma h						
Children 5-17 yrs	% working children	% Child labour	% CL in hazardous work	% CL in Agr	% CL in Ind	% CL in Ser
	0.000501	0.000335	0.000000	0.000000	0.000000	0.000000
	0.000946	0.000631	0.000000	0.000000	0.000000	0.000000
	0.000421	0.000191	0.002234	0.002395	0.000164	0.002229

Similar calculations are carried out for computing the sampling errors of estimates of ratios. These are shown in the preceding page. They involve the same five steps but applied to the linearized variables z_j.

The last two columns of the intermediary calculations use the output values on the sampling errors of the estimated levels to derive the generalized variances as shown in the diagram below[3].

[3] The author is grateful to Alma Kondi, Albania Institute of Statistics (INSTAT) for pointing to an error in the diagram. It has now been corrected.



Generalized variance	
for levels	
σ^2/x^2	$1/x$
0.001664	0.000000
0.004877	0.000002
0.004554	0.000003
0.007986	0.000004
0.137174	0.000093
0.016576	0.000007
0.020444	0.000009
b	a
1450	0.003376
37	0.001293
0.996821	0.003013
1568	5

The top panel calculates the seven sets of regression data corresponding to the seven target variables (number of children 5 to 7 years old; number of working children; number of children engaged in hazardous work; and child labour in agriculture, industry and services).

The results of the regression fit are in the lower panel. The first line gives the estimates of the regression parameters (b=1450 and a=0.003376), and the second line the standard errors of these estimates (σ_b=37 and σ_a = 0.001293, respectively). The third line gives the regression fit (R^2=0.996821) and the sum of squares of the dependent variable (ss_y=0.002013, where $y=\sigma^2/x$). The fourth line gives the F-value of the regression test (F=1568) and the corresponding degrees of freedom (df=5). Finally, the last line gives the sum of squares of the regression fit (ss_{reg}=0.014293) and the residual sum of squares (ss_{res}=0.000045)