RISQ Representativity Indicators for Survey Quality

RISQ manual 2.0

Tools in SAS and R for the computation of R-indicators and partial R-indicators

Vincent de Heij, Barry Schouten
Centraal Bureau voor de Statistiek, The Netherlands

Natalie Shlomo
University of Manchester, United Kingdom

January 31, 2014

Table of contents

Table of contents 1
1. Introduction 1
2. Downloading and installing the RISQ suite 2
3. Getting started 3
3.1 Getting started in R 3
3.2 Getting started in SAS 3
4. The R-indicator 6
4.1 Output in R 7
4.2 Output in SAS 8
5. Bias adjustment and confidence intervals of R-indicators 9
6. Unconditional partial indicators on the variable level 9
6.1 Output in R 10
6.2 Output in SAS 11
7. Unconditional partial indicators within categories 11
7.1 Output in R 12
7.2 Output in SAS 12
8. Conditional partial indicators on the variable level 13
8.1 Output in R 14
8.2 Output in SAS 15
9. Conditional partial indicators within categories 15
9.1 Output in R 16
9.2 Output in SAS 16
10. Bias adjustment and confidence intervals of partial R-indicators 17
11. The coefficient of variation 18
12. General guidelines to R-indicators and partial R-indicators 19
12. Visualising R-indicators in R-cockpit 20
13. Future releases of `RISQ_R-indicators` in SAS and R 20

1. Introduction

This document is one of the two manuals of software developed within project RISQ (Representativity Indicators for Survey Quality). It describes the R and SAS software libraries that can be used for the computation of R-indicators and partial R-indicators. The other manual describes the graphical tool called R-cockpit. The RISQ project is financed by the 7th EU Research Framework Programme. This manual is a second, updated version and includes the various new features that have been added to the R and SAS libraries in RISQ 2.0. The RISQ manual of May 2010 refers to RISQ 1.0.

The RISQ suite is developed in SAS and in R and is available at www.risq-project.eu. In this manual we give basic background to the various indicators developed under the project, we explain how the suite can be used and adapted to any survey data set, and we illustrate its use for the anonymised data set that can be downloaded from the website.

Detailed background to the concepts and ideas behind representativity indicators can be found in the following documents:

- Schouten, B., Cobben, F., Bethlehem, J. (2009), Indicators for the representativeness of survey response, Survey Methodology, 35 (1), 101 – 113.
- Schouten, B., Shlomo, N., Skinner, C. (2011), Indicators for monitoring and improving representativeness of response, Journal of Official Statistics, 27(2), 231 – 253.

- Shlomo, N., Skinner, C., Schouten, B. (2012), Estimation of an indicator of the representativeness of survey response, Journal of Statistical Planning and Inference, 142, 201 – 211.
- Shlomo, N., Schouten, B. (2013), Theoretical properties for partial indicators for representative response, Technical paper, Southampton, University of Southampton, UK
- Shlomo, N., Schouten, B., De Heij, V. (2013), Designing adaptive survey designs using R-indicators, Paper presented at NTTS conference, March 3 – 7, Brussels, Belgium, Available at: http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_63.pdf

Guidelines and a general overview are contained in the following documents:

- Schouten, B., Morren, M., Bethlehem, J., Shlomo, N., Skinner, C. (2009), How to use R-indicators?, RISQ deliverable 3
- Schouten, B., Bethlehem, J. (2009), Representativeness indicators for measuring and enhancing the composition of survey response, RISQ deliverable 9
- Schouten, B., Bethlehem, J., Beulens, K., Kleven, Ø., Loosveldt, G., Rutar, K., Shlomo, N., Skinner, C. (2012), Evaluating, comparing, monitoring and improving representativeness of survey response through R-indicators and partial R-indicators, International Statistical Review, 80 (3), 382 – 399.

Examples of the use of representativity indicators in survey data collection monitoring are given in the following documents:

- Loosveldt, G., Beullens, K. (2009), Fieldwork monitoring, RISQ deliverable 5
- Loosveldt, G., Beullens, K., Luiten, A., Schouten, B. (2010), Improving the fieldwork using R-indicators: applications, RISQ deliverable 6
- Luiten, A., Schouten, B. (2013), Adaptive fieldwork design to increase representative household survey respons. A pilot study in the Survey of Consumer Satisfaction, Journal of Royal Statistical Society, Series A, 176 (1), 169 – 190.
- Schouten, B., Calinescu, M. (2013), Paradata as input to monitoring representativeness and measurement profiles. A case study on the Labour Force Survey, In Improving surveys with paradata (ed. F. Kreuter).

All documents are available at www.risq-project.eu .

2. Downloading and installing the RISQ suite

The SAS and R programs can be downloaded from the RISQ website. From the RISQ website also an anonymised SPSS survey data set can be downloaded. It is called RISQ-test.sav and contains approximately 35,000 persons. In the following we will refer to it as RISQ-test. The data set can be used to test the RISQ suite. It will be used in the examples below.

For the moment a single file contains all the R-code which is needed to determine the R-indicators. In the near future the single file will be replaced by a package. Sourcing the single file will make the functions available in R;

```
> source("RISQ_R-indicators_v2.0.r")
> ls()
[1] "%sub%"                              "getBiasRSampleBased"
[3] "getPartialRConditional"             "getPartialRs"
[5] "getPartialRUnconditional"           "getRIndicator"
[7] "getRSampleBased"                    "getSampleCovTotalPPS"
[9] "getSampleCovTotalSTSI"              "getSampleDesign"
[11] "getSampleStrata"                    "getSampleVarRatio"
[13] "getSampleVarRatioSI"                "getSampleVarTotalPPS"
[15] "getSampleVarTotalSTSI"              "getTrace"
[17] "getVariables"                       "getVariancePartialRConditional"
[19] "getVariancePartialRUnconditional"   "getVarianceRSampleBased"
[21] "weightedVar"
```

Only one function is relevant for a user of the R-code: `getRIndicator`. The user never has to call directly any of the other functions.
In SAS all computations are done within program `RISQ_R-indicators_v2.0.sas`

3. Getting started

3.1 Getting started in R

To load the RISQ-test data set, the function `read.spss` from the package `foreign` is needed. To load the RISQ-test data set `read.spss("RISQ-test.sav")` can be used in the folder where file is stored. To transform the list of vectors which `read.spss` returns, into a data frame[1], the function `as.data.frame` can be used.

```
> library(foreign)
> sampleData <- read.spss("RISQ-test.sav")
> sampleData <- as.data.frame(sampleData)
> summary(sampleData[c("respons", "gender", "age", "urb")])

respons        gender                age                  urb
N/a:    0   Male  :17667   35-39 years: 3572   Very strong:5637
No :16076   Female:17788   40-44 years: 3424   Strong     :9419
Yes:19379                  30-34 years: 3352   Average    :7443
                           50-54 years: 3174   Little     :7864
                           45-49 years: 3106   Not        :5092
                           55-59 years: 2942
                           (Other)    :15885
```

Before the R-indicators can be calculated a response model has to be defined. The left hand side of the formula (the part left of the symbol-~) states the response variable, the right hand side (the part right of the symbol-~) states the model which will be used to fit the response. A model may consist of main effect terms and interaction effect terms. For example, the next three formulas are allowed;

```
> respons ~ gender * age    #  Full model
> respons ~ gender + age    #  Only main effects
> respons ~ gender:age      #  Only interaction effects
```

All variables which are used in the formula have to be members of the data frame with the sample data. The variables on the right hand side of the formula should be factors, the response variable on the left hand side of the formula should either be a factor (logistic regression) or a numeric variable with values zero or one (linear regression). A variable, e.g. age, is transformed into a factor by

```
> sampleData$age <- factor(sampleData$age)
```

A response model can be stored as a formula object but also be fed into the functions directly;

```
> responsModel <- formula(respons ~ gender + age + urb)
```

3.2 Getting started in SAS

The following steps are needed to prepare RISQ-test.sav for computing R-indicators and partial R-indicators and their confidence intervals:

Step 0: Transfer the data set to SAS in SPSS by saving it as a SAS data file.

[1] In R, a data set will usually be an object of the type "data frame". A data frame is usually more convenient than a matrix.

Step 1: The first part of the preparation to run the SAS program is for the user to input information about the dataset, the relevant variables to be used in the logistic regression model and other data set parameters. We refer to the screen shot in figure 3.2.1a and figure 3.2.1b as examples. The first example in figure 3.2.1a does not include interactions in the response model and the second example in figure 3.2.1b includes an interaction.

1. Define the name of the SAS library which contains the dataset and will include the outputs. In figure 3.2.1, the first line of the program defines the **libname** as **RISQ**.
2. Define the following:
   - Size of population – **popsize**
   - Size of sample – **samsize**
   - Number of independent variables in the logistic regression response model (including interactions) – **variablenum.** The names of the variables in the model should be in quotes under **var1, var2,** etc. In the example in figure 3.2.1a, **variablenum=3** and the names of the variables: **var1='agea'; var2='gender'; var3='urb';.** In the example in figure 3.2.1b, **variablenum=2** and the names of the variables are **var1='agea'; var2=gender*urb**;. Note, the variables defining interactions are separated by an asterisk '*'.
   - Number of variables in the logistic regression model that are main effects only – **variablenoint.** In the example in figure 3.2.1a, **variablenoint=3** and in the second example in figure 3.2.1b, **variablenoint=1;**.
   - Number of variables that are used for stratification of the unconditional partial indicator, Pu (see section 6), that are **NOT** used in the logistic regression model – **variablestrat.** The names of the variables should be in quotes under **strat1, strat2**, etc. In the examples in figures 3.2.1a and 3.2.1b, **variablestrat=1** and the name of the variable is **strat1='jobs'**
   - Number of variables that are included in the interactions – **variableinter.** The names of the variables in the interactions should be in quotes under **vvar1, vvar2,** etc. In the example in figure 3.2.1a there are no interactions and **variableinter=0**;. In the example in figure 3.2.1b, **variableinter=2** and the names of the variables: **vvar1='gender'; vvar2='urb';** You should not count the same variable twice, for example, if there were two interactions in the model, eg. **var3='gender*urban';** and **var4='gender*region';,** then **variableinter=3** and **vvar1='gender'; vvar2='urban'; vvar3='region';**
   - The names of all variables are repeated in order to calculate partial R-indicators under the label of **xvar**. We start with the variables defined as main effects listed under **var**, then the variables listed in any interactions under **vvar**, and finally any stratifying variable under **strat.** For example, in figure 3.2.1b where there is an interaction in the response model, we write **xvar1=agea; xvar2=gender; xvar3=urb** and **xvar4=joba**. Note that for these labels, we do not enclose the names in apostrophes.
   - List the number of categories of each variable under the label **nvar** in the same order as they appear under **xvar.** For example, in figure 3.2.1b we write: **nvar1=14; nvar2=2; nvar3=5; nvar4=2;**

Step 2: The second part of the preparation to run the program is to define the labels for the categories of the variables that were defined in step 1 according to the SAS **Proc Format** statement. See figure 3.2.2 for an example of step 2 for the example shown in figure 3.2.1b. In **Proc Format** every variable defined in Step 1 has to be referenced, and for each variable, all its categories have to be stated followed by a label, e.g. **Proc Format; value gender 1="male" 2="female";**. In order to simplify, all variables should be encoded as 1,2,3,4,..etc., and if they are not so defined, they can be transformed in Step 3 below.

Step 3: The last part of the preparation to run the program is to define the dataset, and any necessary transformations or relabeling that may need to be carried out. For instance, the variable **age** in the RISQ-test file was changed to **agea** by collapsing the first three categories as follows: **agex=age-2; if agex=1 or agex=2 then agea=1; if agex ge 3 then agea=agex-1;** In another example, the variable **job** has a value 0-1. We transform this variable to have a value 1-2 as follows: **joba=job+1;**
In addition, the user needs to define a response indicator denoted as **responsesamp1** where **1 is a response and 0 is a non-response**. It's very important that these categories are correctly defined to ensure correct interpretation of partial R-indicators. In the RISQ-test data file, **respons** is the 0-1 indicator for response where 1 is a response and 0 is a non-response.

The user also needs to define the sample design weights, i.e. the inverse of the sample inclusion probabilities, for all sample units (respondents and non-respondents). For simple random sampling, **piinv** is equal to **1/pi** which is the **popsize/samsize** defined in step 1. For any other design, the design weight **d** should be included on the dataset and **piinv** is equal to **d**. See figure 3.2.3 how step 3 is implemented for the RISQ-test file.

Figure 3.2.1a: First part of program RISQ_R-indicators_v2.0.sas - no interaction response model

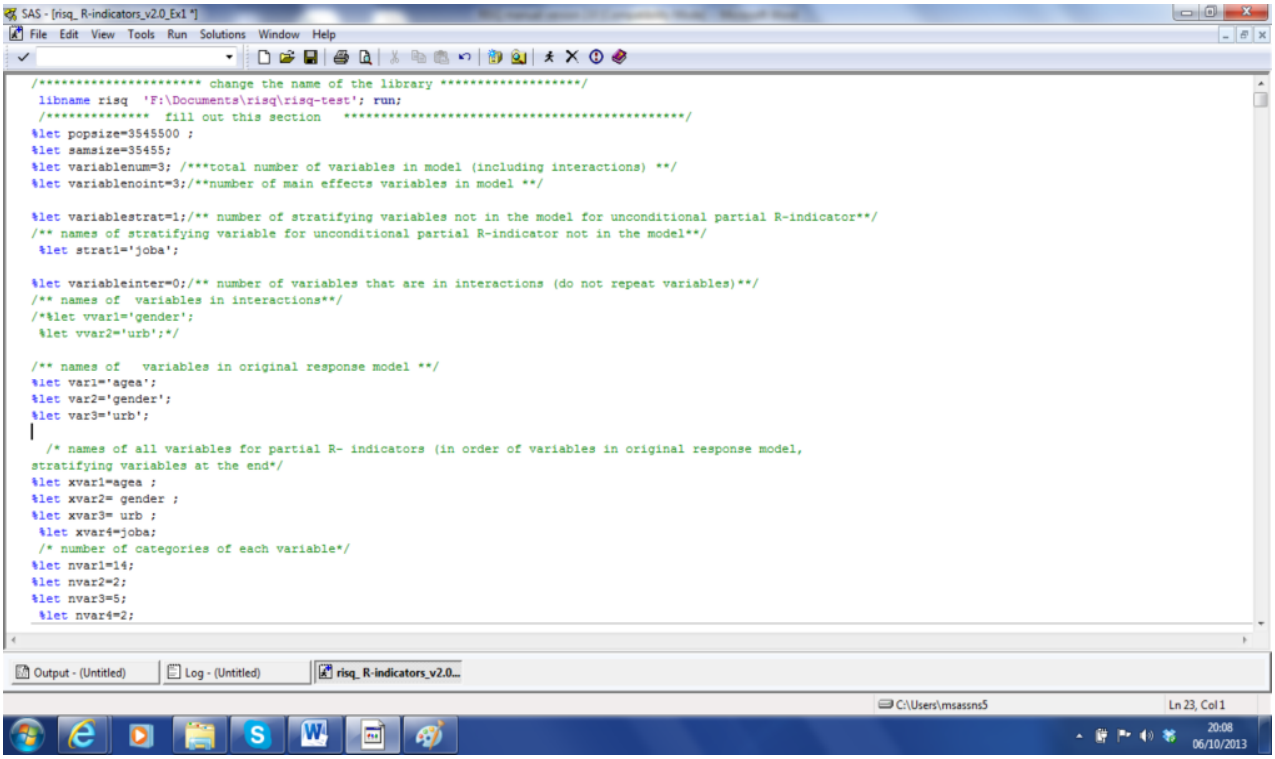


Figure 3.2.1b: First part of program RISQ_R-indicators_v2.0.sas - interaction in response model

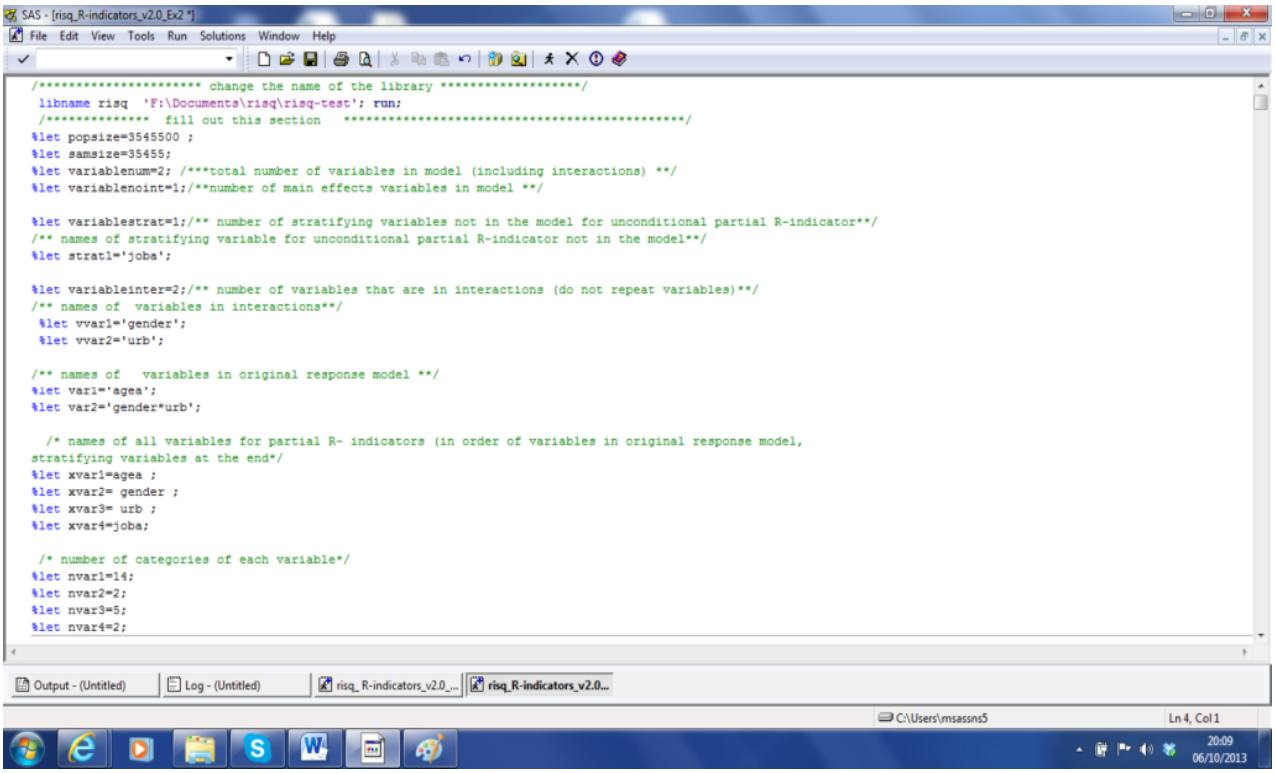


Figure 3.2.2: Labelling the categories of the variables

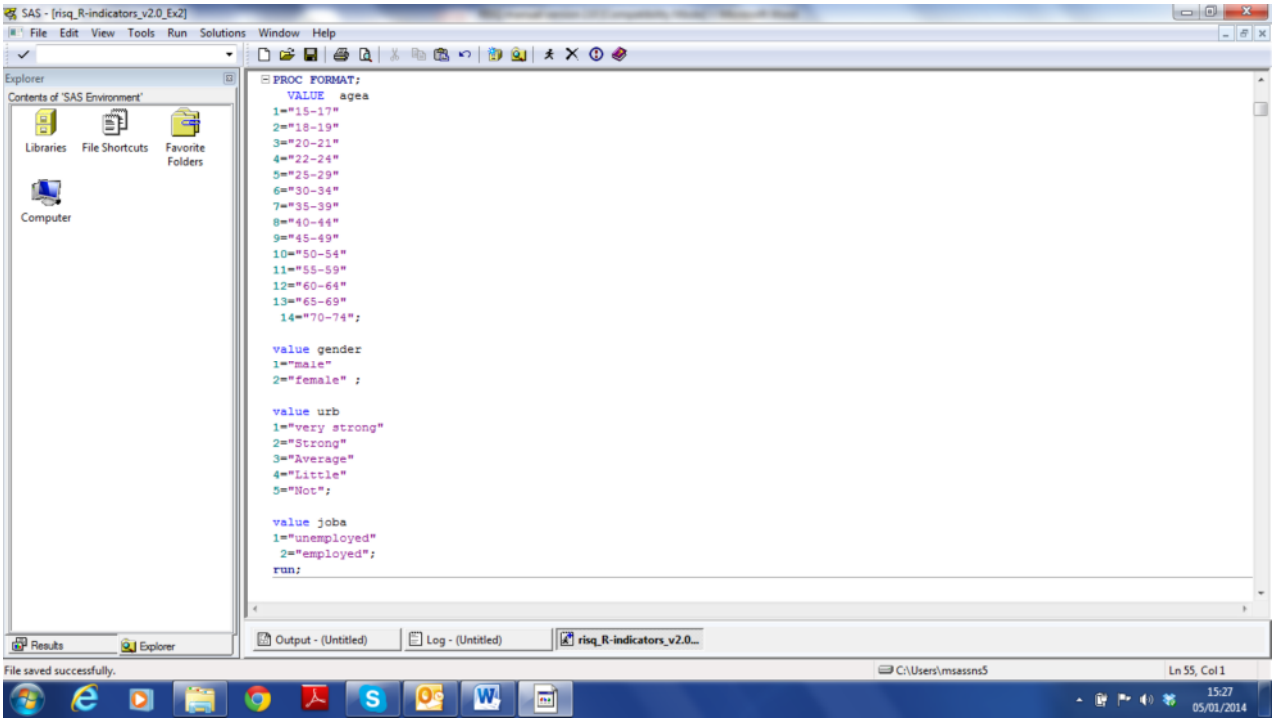
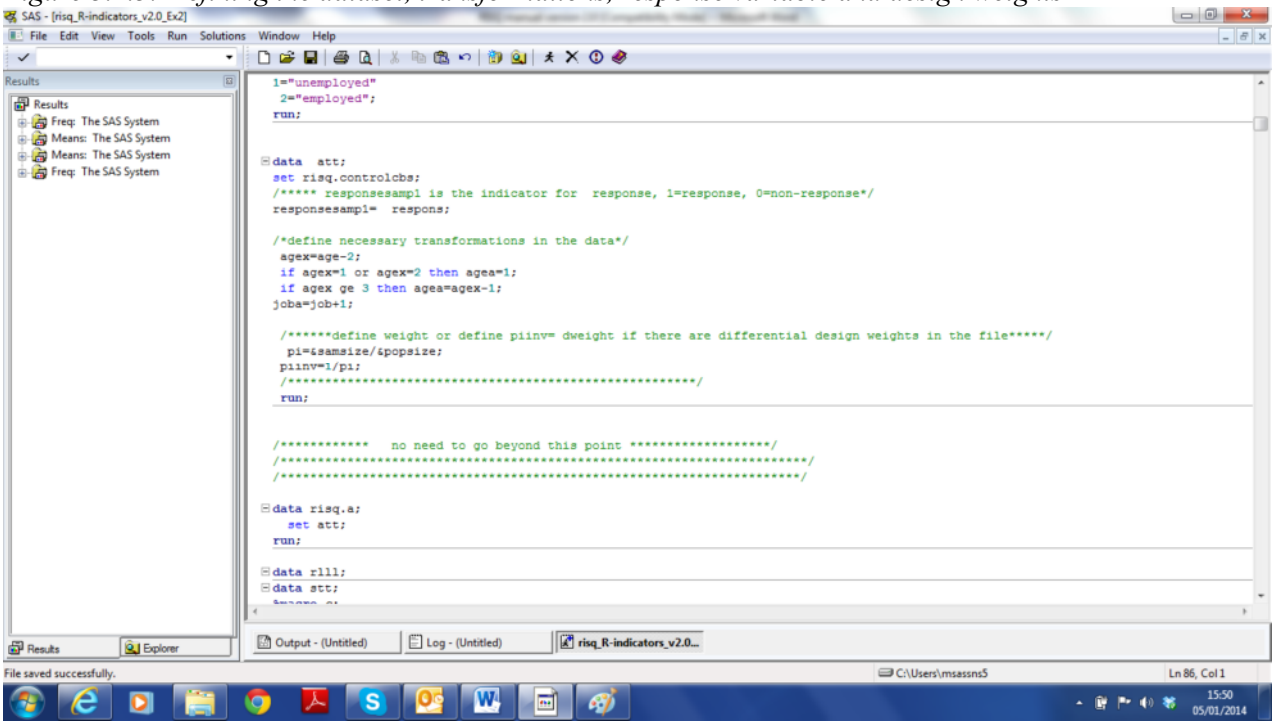


Figure 3.2.3: Defining the dataset, transformations, response variable and design weights



4. The R-indicator

The R-indicator is a transformation of the variance of estimated response propensities to the [0,1] interval. A value equal to 1 implies representative response. A value equal to 0 implies a maximal deviation from representative response.

Suppose the estimated response probabilities for the n elements in the sample are denoted by $\rho_1, \rho_2, \ldots, \rho_n$ and the sample design inclusion weights are denoted by $d_1, d_2, \ldots, d_n$. The design weights are the inverse of the probabilities that a population unit is contained in the survey sample. Then the R-indicator is computed as

$$R = 1 - 2S(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{n} d_i (\rho_i - \bar{\rho})^2}\ , \tag{1}$$

with $\bar{\rho} = \frac{1}{N}\sum_{i=1}^{n} d_i \rho_i$ the weighted sample mean of the estimated response probabilities and N the size of the population.

Response probabilities can be estimated in the R component of the RISQ suite by either a linear or a logistic regression. The default in R is a logistic regression. In SAS response propensities are always estimated by a logistic regression. Let $X = (X_1, X_2, \ldots, X_m)'$ be the vector of independent variables. X needs to be provided by the user. Main effect terms as well as interaction effect terms may be included.

The coefficient of variation is a relevant measure whenever a survey produces estimates for population means and totals only. In those cases it may be used instead of the R-indicator. It is defined as

$$CV = \frac{S(\rho)}{\bar{\rho}}. \tag{2}$$

We return to this measure in section 11.

4.1 Output in R

Once the response model is defined, the R-indicator can be determined;

Option 1:

```
> responsModel <- formula(respons ~ gender + age + urb)
> indicator <- getRIndicator(responsModel,
+     sampleData, sampleWeights, sampleStrata, family)
```

Option 2:

```
> indicator <- getRIndicator(respons ~ gender + age + urb,
+     sampleData, sampleWeights, sampleStrata, family)
```

The response model can either be stored as a formula and then entered as a parameter (option 1) or can be entered directly as a parameter (option 2). The type of link function is `family = 'binomial'` for logistic regression or `family = 'gaussian'` for linear regression. The default is logistic. Properties of the sampling design, the inclusion weights and strata, can be specified by the optional arguments `sampleWeights` and `sampleStrata`. These vectors should have a length equal to the number of rows in the data frame `sampleData`. The type of sampling, simple random sampling (SI), stratified simple random sampling (STSI) or something else, is inferred from the values of `sampleWeights` and `sampleStrata`. If there is only one stratum and all inclusion weights are the same, then SI sampling is assumed. If there is more than one stratum and within each stratum the inclusion weights are the same then STSI sampling is assumed.
The return value of the function `getRIndicator` is a list called `indicator`. The most important components are

R	a bias adjusted estimate for the R-indicator; a bias-adjusted estimate will be determined if the inferred sampling design equals SI or STSI;
RUnadj	an estimate for the R-indicator, without any bias adjustment;
RSE	standard error analytic approximation of the estimated R-indicator !new, standard error is now available for SI and STSI
prop	an estimate for the response propensities;
propMean	the mean of the estimated response propensities which equals the response rate
CV	!new coefficient of variation of response propensities; also referred to as maximal absolute bias
CVSE	!new standard error analytic approximation of the estimated coefficient of variation;

New in the R version of RISQ 2.0 is the estimation of the coefficient of variation and an analytic approximation to its standard error. The coefficient is estimated based on the adjusted variance of response propensities. Furthermore, the standard error approximation for the R-indicator itself is now available also for stratified random sampling. RISQ 1.0 provided standard errors for simple random sampling only.

The components of `indicator` can be assessed by concatenating the name of the component with a "$" to `indicator`. The output is for example

```
> c(indicator$R, indicator$RUnadj, indicator$SE, indicator$propMean)
[1] 0.839748039 0.838129332 0.004107512 0.54658
```

4.2 Output in SAS

Unlike the R suite, this version of SAS does not support stratified sampling. A SAS program for stratified sampling is available on request.

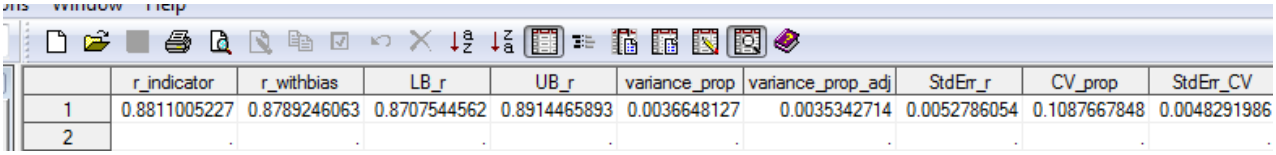
The output for this version of SAS is now ONE dataset which is called **risq.final_output_ex1.** In addition, a CSV file is produced. To change the name and directory of these outputs, these can be changed in the LAST two data step at the very end of the program where you will find the following (the text to be changed is in red):

```
/***** final output file - names and directory of output can be changed here
**********/
data risq.final_output_ex1;
 set u5 ffinal1 ffinal2 ;
 run;
 PROC EXPORT DATA= RISQ.FINAL_OUTPUT_EX1
             OUTFILE= "F:\Documents\risq\risq-test\finaloutputex1.csv"
             DBMS=CSV REPLACE;
     PUTNAMES=YES;
RUN;
```

The first row of the SAS output (and the second row of the CSV output after the labels which appear in the first row) provide the results of the R-indicator and Coefficient of Variation as shown in Figure 4.2.1 for the SAS output.

The **R-indicator** is the adjusted R-indicator value after a bias correction (see section 5), **R_withbias** is the unadjusted R-indicator, **variance_prop** is the original variance of the response propensities (note that response propensities as labelled **rphatsamp** if you are looking through the datasets), **variance_prop_adj** is the bias adjusted variance of the response propensities, **StdErr_r** is the estimated standard error of the R-indicator and **LB_r** and **UB_r** the 95% confidence interval based on a normal approximation. **CV_prop** is the coefficient of variation and its standard error is **StdErr_CV** (see section 11).

Figure 4.2.1: ***SAS Output:*** *R-indicator, standard error and confidence interval, Coefficient of variation and standard error*



	r_indicator	r_withbias	LB_r	UB_r	variance_prop	variance_prop_adj	StdErr_r	CV_prop	StdErr_CV
1	0.8811005227	0.8789246063	0.8707544562	0.8914465893	0.0036648127	0.0035342714	0.0052786054	0.1087667848	0.0048291986
2									

5. Bias adjustment and confidence intervals of R-indicators

R-indicators have a bias that is due to the estimation of response probabilities. In the RISQ suite the bias is approximated analytically. The standard output contains adjusted R-indicator values but unadjusted values are also available.

Suppose the link function h is used in the general linear model for the estimation of the response propensities ρ_i

$$\text{linear regression:} \quad h(x^T\beta) = x^T\beta$$

$$\text{logistic regression:} \quad h(x^T\beta) = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}.$$

Hence, $h(x_i^T\beta)$ is used as a predictor for ρ_i with β a vector that is estimated. Let $\hat{\beta}$ be the estimator and ∇h be the gradient, i.e. the vector with first order derivatives with respect to β.

For simple random samples without replacement, i.e. $d_i = N/n$, the adjusted R-indicator equals

$$R_B = 1 - 2\sqrt{(1+\frac{1}{n}-\frac{1}{N})S^2(\rho) - \frac{1}{n}\sum_{i\in s} z_i^T\left[\sum_{j\in s} z_j x_j^T\right]^{-1} z_i}\,, \tag{3}$$

with $z_i = \nabla h(x_i^T\hat{\beta})x_i$.

Since R-indicators are based on weighted sample variances of estimated probabilities, they also have a standard error and precision. The RISQ suite provides analytic standard error approximations for the R-indicator. The standard errors (c.f. the previous sections on output) can be used to construct confidence intervals. We refer to Shlomo, Skinner and Schouten (2012) for details.

If σ_R is the estimated standard error of the R-indicator, then $[R-\xi_{1-\alpha/2}\sigma_R, R+\xi_{1-\alpha/2}\sigma_R]$ is an $100(1-\alpha)\%$ confidence interval based on a normal approximation. $\xi_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution. The estimated standard error σ_R is **`indicator$RSE`** in R and **StdErr_r** in SAS.

6. Unconditional partial indicators on the variable level

The unconditional partial R-indicator measures the amount of variation of the response probabilities between the categories of a variable. The larger the between-category variation is, the stronger the relationship is and the stronger the impact of the variable on response.

As earlier, let X_k be one of the components of the vector X. Suppose X_k is categorical and has H categories. Let n_h denote the weighted sample size in category h, for $h = 1, 2,..., H$. That means

$$n_h = \sum_{i=1}^{n} d_i \Delta_{h,i} \,, \tag{4}$$

where $\Delta_{h,i}$ is the 0-1 indicator for sample unit i being a member of stratum h. Then $n_1 + n_2 + \ldots + n_H = N$.

Let $\bar{\rho}$ again be the weighted mean response probability in the sample. Furthermore, let $\bar{\rho}_h$ the weighted mean of the response probabilities in category h of X_k.

The unconditional partial indicator for variable X_k is measuring the variation between the response categories of the H categories, and is defined as

$$P_U(X_k) = \sqrt{\frac{1}{N}\sum_{h=1}^{H} n_h \left(\bar{\rho}_h - \bar{\rho}\right)^2} \,. \tag{5}$$

It holds that $P_U(X_k) \leq S(\rho) \leq 0.5$[2]. i.e. the total variation between categories is always smaller than the total variation. The larger the value of (4), the stronger the impact of the variable on nonresponse. By computing and comparing the unconditional partial indicators for a set of variables it can be established for which variables the relationships are strongest.

Also the unconditional partial R-indicators may be subject to bias and like the overall R-indicator they have a standard error. The bias adjustment for the partial R-indicators at the variable level is based on prorating, see Shlomo and Schouten (2013). Based on extensive simulation studies, it was concluded that the bias approximations work satisfactory for sample sizes up to 15,000. For larger surveys it is recommended to use the unadjusted estimates, although they bias adjusted and bias unadjusted estimates are provided both in the output. New in RISQ 2.0 is an analytic approximation to the standard error of the unconditional partial R-indicator. The approximated standard error is taken equal to the standard error of the standard deviation of the estimated response propensities as if the response model consists only of the selected variable X_k. We refer to Shlomo, Schouten and De Heij (2013) for details.

6.1 Output in R

To determine unconditional partial indicators, the optional argument `withPartials` of the function `getRIndicator` should be set to `TRUE`;

```
> indicator <- getRIndicator(responsModel, sampleData,
+      sampleWeights,
+      sampleStrata,
+      withPartials = TRUE)
```

Just as earlier, the return value `indicator` of the function `getRIndicators` contains a component `partials` containing the estimates for the partial R-indicators. The component `partials$byVariables` of the list `indicator` is a data frame with the unconditional and conditional partial indicators for each variable in the model. The data frame contains the following columns:

`variable`	the name of the variable;
`Pu`	a bias adjusted estimate for the unconditional, partial indicator;
`PuUnadj`	an estimate for the partial unconditional, indicator, without any bias adjustment;
`PuSE`	`!new` standard error analytic approximation of estimated unconditional partial indicator

[2] $S(\rho)$ attains its maximum value when half of the ρ_i's are 0 and the rest are 1.

The data frame `partials$byVariables` are found by

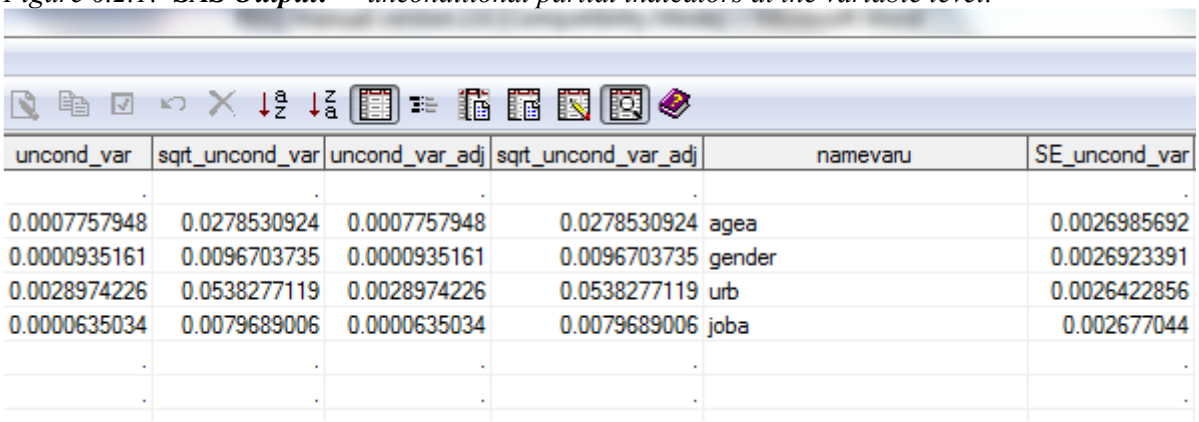
```
> indicator$partials$byVariables
  variable         Pu     PuUnadj        PuSE         Pc      PcUnadj  PcSEApprox
1   gender 0.00949666 0.00967045 0.002692305 0.00909051 0.009256866 0.002692305
2      age 0.02735313 0.02785369 0.002698560 0.02540499 0.025869904 0.002698560
3      urb 0.05286051 0.05382786 0.002642265 0.05201269 0.052964532 0.002642265
```

which contains both unconditional and conditional partial R-indicators. We return to conditional partial R-indicators in section 8.

6.2 Output in SAS

The unconditional variable level partial R-indicators appear in the single SAS and CSV file in the 10th column starting in the second row (or third row of the CSV file). For the example on the test data with no interaction as shown in Figure 3.2.1a, we obtain the results shown in Figure 6.2.1.

Figure 6.2.1: ***SAS Output:*** *- unconditional partial indicators at the variable level.*



uncond_var	sqrt_uncond_var	uncond_var_adj	sqrt_uncond_var_adj	namevaru	SE_uncond_var
.	.	.	.		.
0.0007757948	0.0278530924	0.0007757948	0.0278530924	agea	0.0026985692
0.0000935161	0.0096703735	0.0000935161	0.0096703735	gender	0.0026923391
0.0028974226	0.0538277119	0.0028974226	0.0538277119	urb	0.0026422856
0.0000635034	0.0079689006	0.0000635034	0.0079689006	joba	0.002677044
.	.	.	.		.
.	.	.	.		.

Note that the size of the dataset is over 15,000 and hence there is no bias correction at the variable level partial R-indicator. The **uncond_var** is the unadjusted squared unconditional variable level partial R-indicator and **uncond_var_adj** is with the bias correction when the procedure is carried out for smaller sample sizes. **sqrt_uncond_var** is the unconditional variable level partial R-indicator and **sqrt_uncond_var_adj** is with the bias correction. The standard error of the unconditional variable level partial R-indicator is called **SE_uncond_var**.

7. Unconditional partial indicators within categories

The unconditional partial R-indicator can give more information about the relationship of a variable X_k and response behaviour if this indicator is computed for each category of X_k separately. It is clear from (4) that each category h contributes an amount

$$\frac{n_h}{n}(\bar{\rho}_h - \bar{\rho})^2 \tag{6}$$

to $P_U(X_k)$. The unconditional partial indicators within categories are obtained by taking the square root of the quantities in (6), giving

$$P_U(X_k, h) = \sqrt{\frac{n_h}{N}}(\bar{\rho}_h - \bar{\rho}). \tag{7}$$

$P_U(X_k, h)$ can assume positive and negative values. A positive value means that the particular category is over-represented. A negative value means that the particular category is under-represented.

For the category level the bias adjustment of the partial R-indicators is removed in RISQ 2.0. Based on a simulation study, Shlomo and Schouten (2013) recommend to not perform any bias adjustment at the category level. In RISQ 2.0, an analytic approximation to the standard error is added, following Shlomo, Schouten and De Heij (2013).

7.1 Output in R

The component `partials$byCategories` is a list, containing the partial indicators within categories for each variable in the model. Each component in the list `partials$byCategories` is a data frame with the unconditional and conditional partial indicators within categories of a variable.
Each component of `partials$byCategories` is a data frame whose name equals the name of the variable. One example is `indicator$partials$byCategories$gender`. Most of the columns in the data frame equal the columns in the data frame `indicator$partials$byVariables`. The column `variable` is replaced by the column `category` containing the names of the categories.

```
> indicator$partials$byCategories
$gender
  category      PuUnadj    PuUnadjSE     PcUnadj    PcUnadjSE
1   Female  0.006826362 0.001464660 0.006539714 0.001889557
2     Male -0.006849699 0.001469667 0.006551467 0.001893023

$age
             category       PuUnadj    PuUnadjSE      PcUnadj    PcUnadjSE
1          0-17 years -9.671122e-03 0.002504006 0.0101408961 0.002630584
2         18,19 years  2.796507e-03 0.002875602 0.0026315899 0.003136714
3         20-24 years -6.474036e-03 0.002500824 0.0045119749 0.002724876
4         25-29 years -1.355544e-02 0.002374886 0.0111171122 0.002568662
5         30-34 years -3.498266e-03 0.002476968 0.0025213687 0.003045249
6         35-39 years  2.720500e-03 0.002542023 0.0030693711 0.002855867
7         40-44 years  4.624138e-05 0.002519602 0.0003572775 0.012434739
8         45-49 years  4.985914e-03 0.002630292 0.0043140078 0.002700415
9         50-54 years -2.813430e-03 0.002502324 0.0040327589 0.002738852
10        55-59 years -2.255802e-03 0.002534361 0.0035377203 0.002814882
11        60-64 years  7.004059e-03 0.002781496 0.0060849785 0.002634701
12        65-69 years  8.283321e-03 0.002870593 0.0075442962 0.002608567
13        70-74 years  1.654195e-02 0.003117646 0.0160690303 0.002526770
14 75 years and older  5.819584e-05 0.002614814 0.0002973371 0.015193023

$urb
     category      PuUnadj    PuUnadjSE     PcUnadj    PcUnadjSE
1     Average  0.010083497 0.002192683 0.010067456 0.002328767
2      Little  0.016929938 0.002200976 0.016460659 0.002309831
3         Not  0.018071340 0.002532884 0.017934496 0.002419178
4      Strong -0.001599985 0.001941088 0.002560629 0.001879914
5 Very strong -0.046690533 0.001817152 0.045877355 0.002420162
```

7.2 Output in SAS

The unconditional categorical level partial R-indicators appear in the single SAS and CSV file in the appropriate column starting in the sixth row (or seventh row of the CSV file). For the example on the test data with no interaction as shown in Figure 3.2.1a , we obtain the results shown in Figure 7.2.1.

Figure 7.2.1: ***SAS Output:*** *- all unconditional partial indicators at the category level.*

agea	popsize	avg_propensity_cat	avg_propensity	uncond_cat	sqrt_uncond_cat	gender	urb	joba	SE_uncond_cat
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
1	141900	0.4982381951	0.5465799024	0.0000935296	-0.009671068	.	.	.	0.002506514
2	101400	0.5631162455	0.5465799024	7.8205881E-6	0.0027965314	.	.	.	0.0028784812
3	258700	0.5226130439	0.5465799024	0.0000419123	-0.006473969	.	.	.	0.0025033283
4	264800	0.4969788477	0.5465799024	0.0001837479	-0.013555363	.	.	.	0.0023772644
5	335200	0.5352028309	0.5465799024	0.0000122374	-0.003498193	.	.	.	0.0024794485
6	357200	0.5551510961	0.5465799024	7.4014461E-6	0.0027205599	.	.	.	0.0025445685
7	342400	0.5467289255	0.5465799024	2.1446847E-9	0.0000463107	.	.	.	0.0025221248
8	310600	0.5634255133	0.5465799024	0.0000248598	0.0049859595	.	.	.	0.0026329254
9	317400	0.537177037	0.5465799024	7.914981E-6	-0.002813358	.	.	.	0.0025048303
10	294200	0.5387491214	0.5465799024	5.0883308E-6	-0.002255733	.	.	.	0.0025368988
11	224400	0.5744204614	0.5465799024	0.000049057	0.0070040725	.	.	.	0.0027842806
12	184600	0.582881533	0.5465799024	0.000068613	0.008283297	.	.	.	0.0028734656
13	155200	0.6256398429	0.5465799024	0.000273607	0.0165410699	.	.	.	0.0031207465
14	257500	0.5467960598	0.5465799024	3.3934376E-9	0.0000582532	.	.	.	0.002617432
.	1766700	0.5368764696	0.5465799024	0.0000469176	-0.006849645	1	.	.	0.0014711382
.	1778800	0.5562173292	0.5465799024	0.0000465985	0.0068263084	2	.	.	0.0014661261
.	563700	0.4294837697	0.5465799024	0.0021799958	-0.046690425	.	1	.	0.0018189716
.	941900	0.5434757752	0.5465799024	2.5598017E-6	-0.001599938	.	2	.	0.0019430314
.	744300	0.5685876883	0.5465799024	0.0001016771	0.0100835073	.	3	.	0.0021948781
.	786400	0.58252758	0.5465799024	0.0002866208	0.0169298787	.	4	.	0.0022031791
.	509200	0.5942649352	0.5465799024	0.0003265691	0.0180712227	.	5	.	0.002535418
.	1753200	0.5546371745	0.5465799024	0.0000321018	0.0056658493	.	.	1	0.0014778567
.	1792300	0.5386984041	0.5465799024	0.0000314015	-0.005603707	.	.	2	0.0014616476

The estimated size of the population is in the variable **popsize**, the average of the propensity score for the category is in **avg_propensity_cat** and the overall average propensity is in **avg_propensity**. The squared unconditional category level partial R-indicator is in **uncond_cat** and the unconditional category level partial R-indicator is in **sqrt_uncond_cat**. The standard error is in **SE_uncond_cat.**

8. Conditional partial indicators on the variable level

Conditional partial indicators can only be computed for variables that are included in the response model. These indicators measure the relative importance of a variable, i.e. the impact of a variable conditional on all other variables in the response model. As such conditional partial R-indicators attempt to isolate the part of the deviation of representative response that is attributable to a variable alone.

The conditional partial indicator for a variable X_k is obtained by cross-classification of all model variables, but with the exception of X_k itself. Suppose, this cross-classification results in L cells U_1, U_2, …, U_L. Let n_l denote the weighted sample size in cell l, for $l = 1, 2, .., L$. Then again $n_1 + n_2 + \ldots + n_L = N$. Furthermore, let $\bar{\rho}_l$ the mean of the response probabilities in cell l.

The conditional partial indicator for variable X_k is now defined as

$$P_C(X_k) = \sqrt{\frac{1}{N}\sum_{l=1}^{L}\sum_{i\in U_l} d_i\left(\rho_i - \bar{\rho}_l\right)^2} \quad . \tag{8}$$

To say it in words: $P_C(X_k)$ is the remaining within cell variation of the response probabilities if the variable X_k is removed from the cross-classification. If, on the one hand, the remaining variation is large,

this can apparently not be accounted for by the other variables. So, there is an important role for X_k. If, on the other hand, the remaining variation is small, the other variables are capable of explaining the variation. It can be concluded that there need not be a role for X_k in reducing the lack of representativity.

Also here it can be remarked that $P_C(X_k) \leq S(\rho) \leq 0.5$, i.e. the total variation within categories is smaller than the total variation, and again a larger value for $P_C(X_k)$ implies a stronger conditional impact.

The conditional partial R-indicators may also be subject to bias and they have a standard error. In RISQ 2.0, the bias adjustment for the partial R-indicators at the variable level is left unchanged and is based on prorating, see Shlomo and Schouten (2013). Based on simulation studies, it is again recommended to use the adjusted estimates for sample sizes smaller than 15,000 and the unadjusted estimates for larger sample sizes. Both estimates are, however, provided. New in RISQ 2.0 is an analytic approximation to the standard error of the conditional partial R-indicator. The approximation in SAS and R is different. In SAS, the approximated standard error is taken to be equal to the standard error of the standard deviation of the estimated response propensities as if the response model consists only of all other variable X_k^- and not including the selected variable X_k. Based on simulation studies it was concluded that this approximation works satisfactory under most circumstances but may produce invalid results when the R-indicator attains values close to one. For this reason, the R code uses a conservative approximation, namely to take the standard error approximation of the unconditional variable-level partial R-indicator, which is always larger. We refer to Shlomo, Schouten and De Heij (2013) for details.

8.1 Output in R

To determine conditional partial indicators, the optional argument `withPartials` of the function `getRIndicator` should again be set to `TRUE`;

```
> indicator <- getRIndicator(responsModel, sampleData,
+     sampleWeights,
+     sampleStrata,
+     withPartials = TRUE)
```

The return value of the function `getRIndicators` contains a component `partials` containing the estimates for the partial R-indicators. The component `partials$byVariables` of the list `indicator` is a data frame with the unconditional and conditional partial indicators for each variable in the model. The data frame contains the following columns:

`variable`	the name of the variable;
`Pc`	a bias adjusted estimate for the conditional partial indicator; a bias-adjusted estinmate will be determined if the inferred sampling design equals SI or STSI;
`PcUnadj`	an estimate for the partial conditional indicator, without any bias adjustment.
`PcSEApprox`	`!new` standard error analytic approximation of estimated conditional partial indicator; equals `PuSE`

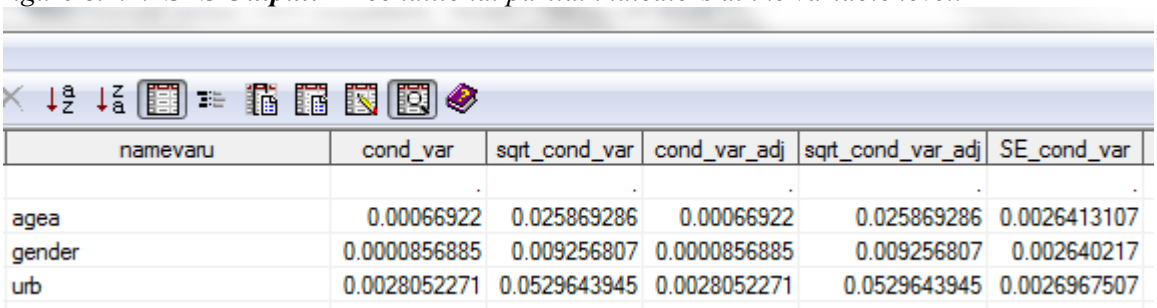
The output is

```
> indicator$partials$byVariables
  variable         Pu    PuUnadj        PuSE         Pc     PcUnadj  PcSEApprox
1   gender 0.00949666 0.00967045 0.002692305 0.00909051 0.009256866 0.002692305
2      age 0.02735313 0.02785369 0.002698560 0.02540499 0.025869904 0.002698560
3      urb 0.05286051 0.05382786 0.002642265 0.05201269 0.052964532 0.002642265
```

8.2 Output in SAS

The conditional variable level partial R-indicators appear in the single SAS and CSV file in the appropriate column starting in the second row (or third row of the CSV file). For the example on the test data with no interaction as shown in Figure 3.2.1a , we obtain the results shown in Figure 8.2.1.

Figure 8.2.1: ***SAS Output:*** *- conditional partial indicators at the variable level.*



namevaru	cond_var	sqrt_cond_var	cond_var_adj	sqrt_cond_var_adj	SE_cond_var
	.	.	.	.	.
agea	0.00066922	0.025869286	0.00066922	0.025869286	0.0026413107
gender	0.0000856885	0.009256807	0.0000856885	0.009256807	0.002640217
urb	0.0028052271	0.0529643945	0.0028052271	0.0529643945	0.0026967507

Note that the size of the dataset is over 15,000 and hence there is no bias correction at the variable level partial R-indicator. The **cond_var** is the unadjusted squared conditional variable level partial R-indicator and **cond_var_adj** is with the bias correction when the procedure is carried out for smaller sample sizes. **sqrt_cond_var** is the conditional variable level partial R-indicator and **sqrt_cond_var_adj** is with the bias correction. The standard error of the conditional variable level partial R-indicator is called **SE_cond_var**.

9. Conditional partial indicators within categories

The conditional partial indicators can give even more insight when they are computed for each category of a variable separately. The remaining within cell variation of the response probabilities after removing a variable X_k from the cross-classification, is computed for each category of X_k separately. Let again X_k have H categories, labelled $h=1,2,...,H$, and $\Delta_{h,i}$ be the 0-1 indicator for category h. From (7) it can be seen that each category h contributes an amount

$$\frac{1}{N}\sum_{l=1}^{L}\sum_{i\in U_l} d_i \Delta_{h,i}\left(\rho_i - \bar{\rho}_l\right)^2 \tag{9}$$

to $P_C(X_k)$. The conditional partial indicators within categories are then obtained by taking the square root of (9)

$$P_C(X_k,h) = \sqrt{\frac{1}{N}\sum_{l=1}^{L}\sum_{i\in U_l} d_i \Delta_{h,i}\left(\rho_i - \bar{\rho}_l\right)^2}\ . \tag{10}$$

The category-level conditional partial R-indicators are always larger than or equal to zero. A large value of (10) does not correspond to either under- or over-representation. Such an interpretation cannot be given as within some cells l the category may be over-represented while in other cells it may be under-represented. Hence, the subpopulation corresponding to a category may be overrepresented in some cells and underrepresented in others. The conditional partial indicator within a category $P_C(X_k,h)$ must be interpreted as the impact of that category on the deviation from representative response after conditioning on the other variables. The larger the indicator the larger the impact of that category and the more interesting the corresponding subpopulation becomes in nonresponse reduction methods.

Also for the category level conditional partial R-indicator the bias adjustment is removed in RISQ 2.0. This change is based on the same simulation study described in Shlomo and Schouten (2013). In RISQ 2.0, an analytic approximation to the standard error is added, following Shlomo, Schouten and De Heij (2013).

9.1 Output in R

As we did for the unconditional partial indicator at the category level, we will consider the data frame `partials$byCategories`, but this time we focus on the last two columns of the data frame: `Pc` and `PcUnadj`. The component `partials$byCategories` is a list, containing the partial indicators within categories for each variable in the model. Each component of `partials$byCategories` is a data frame whose name equals the name of the variable. One example is `indicator$partials$byCategories$gender`. Most of the columns in the data frame equal the columns in the data frame `indicator$partials$byVariables`. The column `variable` is replaced by the column `category` containing the names of the categories.

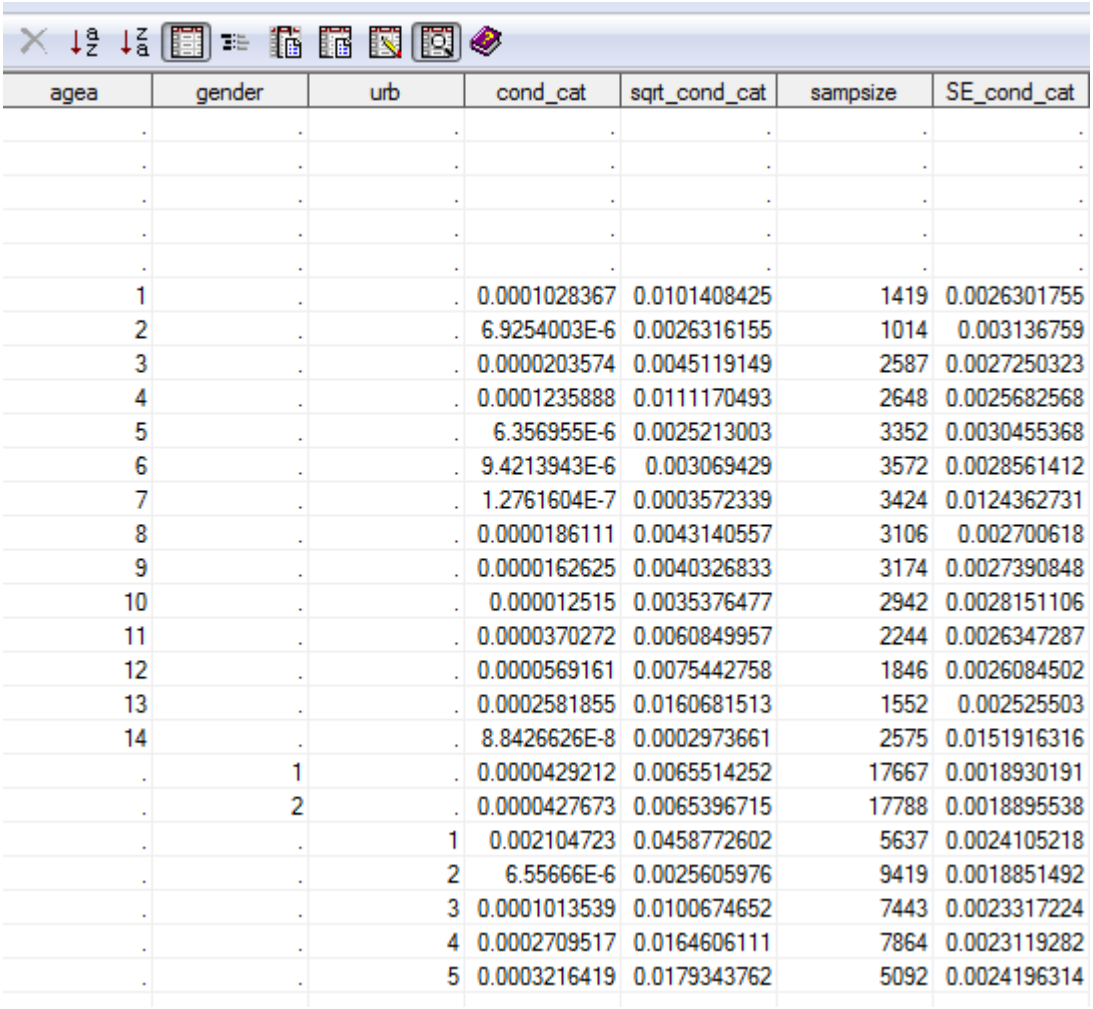
```
> indicator$partials$byCategories
$gender
  category      PuUnadj    PuUnadjSE     PcUnadj   PcUnadjSE
1   Female  0.006826362 0.001464660 0.006539714 0.001889557
2     Male -0.006849699 0.001469667 0.006551467 0.001893023

$age
             category       PuUnadj    PuUnadjSE      PcUnadj    PcUnadjSE
1          0-17 years -9.671122e-03 0.002504006 0.0101408961 0.002630584
2         18,19 years  2.796507e-03 0.002875602 0.0026315899 0.003136714
3         20-24 years -6.474036e-03 0.002500824 0.0045119749 0.002724876
4         25-29 years -1.355544e-02 0.002374886 0.0111171122 0.002568662
5         30-34 years -3.498266e-03 0.002476968 0.0025213687 0.003045249
6         35-39 years  2.720500e-03 0.002542023 0.0030693711 0.002855867
7         40-44 years  4.624138e-05 0.002519602 0.0003572775 0.012434739
8         45-49 years  4.985914e-03 0.002630292 0.0043140078 0.002700415
9         50-54 years -2.813430e-03 0.002502324 0.0040327589 0.002738852
10        55-59 years -2.255802e-03 0.002534361 0.0035377203 0.002814882
11        60-64 years  7.004059e-03 0.002781496 0.0060849785 0.002634701
12        65-69 years  8.283321e-03 0.002870593 0.0075442962 0.002608567
13        70-74 years  1.654195e-02 0.003117646 0.0160690303 0.002526770
14 75 years and older  5.819584e-05 0.002614814 0.0002973371 0.015193023

$urb
     category       PuUnadj    PuUnadjSE     PcUnadj    PcUnadjSE
1     Average  0.010083497 0.002192683 0.010067456 0.002328767
2      Little  0.016929938 0.002200976 0.016460659 0.002309831
3         Not  0.018071340 0.002532884 0.017934496 0.002419178
4      Strong -0.001599985 0.001941088 0.002560629 0.001879914
5 Very strong -0.046690533 0.001817152 0.045877355 0.002420162
```

9.2 Output in SAS

The conditional categorical level partial R-indicators appear in the single SAS and CSV file in the appropriate column starting in the sixth row (or seventh row of the CSV file). For the example on the test data with no interaction as shown in Figure 3.2.1a , we obtain the results shown in Figure 9.2.1.

The sample size is in the variable **sampsize**. The squared conditional category level partial R-indicator is in **cond_cat** and the conditional category level partial R-indicator is in **sqrt_cond_cat**. The standard error is in **SE_cond_cat.**

Figure 9.2.1: ***SAS Output:*** *- all conditional partial indicators at the category level.*



agea	gender	urb	cond_cat	sqrt_cond_cat	sampsize	SE_cond_cat
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1	.	.	0.0001028367	0.0101408425	1419	0.0026301755
2	.	.	6.9254003E-6	0.0026316155	1014	0.003136759
3	.	.	0.0000203574	0.0045119149	2587	0.0027250323
4	.	.	0.0001235888	0.0111170493	2648	0.0025682568
5	.	.	6.356955E-6	0.0025213003	3352	0.0030455368
6	.	.	9.4213943E-6	0.003069429	3572	0.0028561412
7	.	.	1.2761604E-7	0.0003572339	3424	0.0124362731
8	.	.	0.0000186111	0.0043140557	3106	0.002700618
9	.	.	0.0000162625	0.0040326833	3174	0.0027390848
10	.	.	0.000012515	0.0035376477	2942	0.0028151106
11	.	.	0.0000370272	0.0060849957	2244	0.0026347287
12	.	.	0.0000569161	0.0075442758	1846	0.0026084502
13	.	.	0.0002581855	0.0160681513	1552	0.002525503
14	.	.	8.8426626E-8	0.0002973661	2575	0.0151916316
.	1	.	0.0000429212	0.0065514252	17667	0.0018930191
.	2	.	0.0000427673	0.0065396715	17788	0.0018895538
.	.	1	0.002104723	0.0458772602	5637	0.0024105218
.	.	2	6.55666E-6	0.0025605976	9419	0.0018851492
.	.	3	0.0001013539	0.0100674652	7443	0.0023317224
.	.	4	0.0002709517	0.0164606111	7864	0.0023119282
.	.	5	0.0003216419	0.0179343762	5092	0.0024196314

10. Bias adjustment and confidence intervals of partial R-indicators

As for the R-indicators, partial R-indicators have a bias and standard error.

In the RISQ suite the bias of the variable-level partial R-indicators is adjusted by prorating the overall R-indicator bias over the partial R-indicators. That means that the estimated bias of the variance of response probabilities $B(S^2(\rho))$ is multiplied by the ratio between the square of the partial R-indicator and $S^2(\rho)$. This approximation is motivated by the fact that the partial R-indicators are between and within variances which are components of the total variance of response probabilities $S^2(\rho)$. The resulting, prorated bias is then subtracted from the between variance (unconditional partial R-indicators) or the within variance (conditional partial R-indicators). And the partial R-indicators are computed by taking the square root of the adjusted between or within variance.

Let $S^2_{W,unadj}(\rho)$ and $S^2_{B,unadj}(\rho)$ denote, respectively, the unadjusted within variance and the unadjusted between variance of the estimated response propensities. Both variance terms are adjusted for bias in the following way

$$S^2_W(\rho) = S^2_{W,unadj}(\rho) - B(S^2(\rho))\frac{S^2_{W,unadj}(\rho)}{S^2(\rho)} \qquad (11)$$

$$S_B^2(\rho) = S_{W,unadj}^2(\rho) - B(S^2(\rho))\frac{S_{B,unadj}^2(\rho)}{S^2(\rho)} \tag{12}$$

and the adjusted partial R-indicators at the variable level are computed by taking square roots.

The category-level partial R-indicators are not adjusted for bias following recommendations in Shlomo and Schouten (2013).

For details about the standard error approximations for both variable-level and category-level partial R-indicators we refer to Shlomo, Schouten and De Heij (2013). Here, we restrict ourselves to a summary:

- The standard error for the variable-level unconditional partial R-indicator is approximated by the standard error for the standard deviation of the estimated response propensities restricted to a model with only the selected variable. See Shlomo, Skinner and Schouten (2012) for details.
- The standard error for the variable-level conditional partial R-indicator approximated by the standard error for the standard deviation of the estimated response propensities restricted to a model with all variables except the selected variable. See Shlomo, Skinner and Schouten (2012) for details. This approximation does not behave well under all circumstances. For this reason in R the conservative choice is made to use the standard error approximation for the unconditional partial R-indicator at the variable-level.
- The standard error for the category-level unconditional partial R-indicator follows the approximation in Shlomo, Schouten and De Heij (2013).
- The standard error for the category-level conditional partial R-indicator follows the approximation in Shlomo, Schouten and De Heij (2013).

11. The coefficient of variation

In all RISQ deliverables, the R-indicators are interpreted in terms of the impact of nonresponse on survey estimation by considering the standardized bias of the design-weighted response mean $\hat{\bar{y}}_r$ of a survey variable y

$$\frac{|B(\hat{\bar{y}}_r)|}{S(y)} = \frac{|Cov(y,\rho_Y)|}{\bar{\rho}S(y)} = \frac{|Cov(y,\rho_{\aleph})|}{\bar{\rho}S(y)} \leq \frac{S(\rho_{\aleph})}{\bar{\rho}} = \frac{1-R(\aleph)}{2\bar{\rho}}, \tag{13}$$

with $\bar{\rho}$ the average response propensity and $\aleph$ the vector of auxiliary variables explaining response behaviour. The vector $\aleph$ is unknown and, as a consequence, we do not know $\rho_{\aleph}$. Since we are interested in the general representativeness of a survey, i.e. not the representativeness with respect to single survey items, we use as an approximation for (13)

$$CV(X) = \frac{1-R(\rho_X)}{2\bar{\rho}}. \tag{14}$$

CV is the coefficient of variation of the estimated response propensities and represents the maximal absolute standardized bias under the scenario that non-response correlates maximally to the selected auxiliary variables. ρ_X are the response propensities with a response model based on X. The coefficient of variation (14) is estimated by

$$CV(\hat{\rho}_X) = \frac{S(\hat{\rho}_X)}{\hat{\bar{\rho}}}. \tag{15}$$

The standard error of (15) is derived using the approximation

$$Var(CV(\hat{\rho}_X)) \cong \frac{S^2(\hat{\rho}_X)}{\hat{\bar{\rho}}^2}\left[\frac{Var(\hat{\bar{\rho}})}{\hat{\bar{\rho}}^2} + \frac{Var(S(\hat{\rho}_X))}{S(\hat{\rho}_X)^2} - 2\frac{Cov(\hat{\bar{\rho}}, S(\hat{\rho}_X))}{\hat{\bar{\rho}} S(\hat{\rho}_X)}\right]. \qquad (16)$$

Let the variance of the standard deviation of response propensities be denoted by S^2. It can be reasoned that the covariance between de mean and standard deviation of the response propensities in (16), $Cov(\hat{\bar{\rho}}, S(\hat{\rho}_X))$, is negligible as long as $\hat{\bar{\rho}}$ is roughly in the range $[0.2,0.8]$. In the extreme case where all response propensities are either zero or one, $S(\hat{\rho}_X)$ is approximately equal to $S(\hat{\rho}_X) = \sqrt{\hat{\bar{\rho}}(1-\hat{\bar{\rho}})}$. For $\hat{\bar{\rho}} \in [0.2,0.8]$ this function is very flat and covariances must be small. For values of $\hat{\bar{\rho}} \leq 0.2$, there is a positive covariance, and for $\hat{\bar{\rho}} \geq 0.8$ there is a negative covariance. Since it can be expected that response propensities will not all be zero or one, even for values outside the range $[0.2,0.8]$, the covariance is expected to be small. The variance of the average response propensity, $Var(\hat{\bar{\rho}})$, is also small. It can be approximated by $S^2(\hat{\rho}_X)/n$, with n the sample size.

Given these considerations, the approximation (15) is rewritten to

$$Var(CV(\hat{\rho}_X)) \cong \frac{S^2(\hat{\rho}_X)}{\hat{\bar{\rho}}^2}\left[\frac{S^2(\hat{\rho}_X)}{n\hat{\bar{\rho}}^2} + \frac{S^2}{S(\hat{\rho}_X)^2}\right] = \frac{S^2}{\hat{\bar{\rho}}^2} + \frac{S^4(\hat{\rho}_X)}{n\hat{\bar{\rho}}^4}. \qquad (16)$$

CV is referred to as the maximal bias or coefficient of variation. In RISQ 2.0 it is now available and is computed along with the R-indicator and response rate. The analytic standard error approximation given by (16) is also available;

```
> indicator$CV
[1] 0.1107595
> indicator$CVSE
[1] 0.004806956
```

In SAS the coefficient of variation is not implemented. It can, however, be derived simply by using (14). The standard error approximation cannot be derived as quickly and would need additional programming using (16).

12. General guidelines to R-indicators and partial R-indicators

The following recommendations must be kept in mind when using the R-indicators and partial R-indicators:

- R-indicators and partial R-indicators cannot be evaluated or presented separately from the variables X that were used in the response model and should always be presented together with X.
- When comparing different surveys, one should use the same model for nonresponse, where the variables X, have the same categories.
- R-indicators should be adjoined by a confidence interval in order to indicate the uncertainty due to the estimation based on a sample.
- The inclusion of response-unrelated variables into the response model leads to an increase of the standard errors of R-indicators. It is recommendable to restrict analysis to variables X for which it is known from the literature that they relate to response behaviour.
- R-indicators measure the distance to a fully representative response; they do not reflect the impact of non-response on the bias of (weighted) means or the contrast of survey variables, and nor does the

response rate. The coefficient of variation combines the response rate and the R-indicator and is designed to make comparisons of non-response bias under worst case scenarios.

The various indicators may be used to compare different surveys or a single survey in time. When comparing different surveys, we recommend to fix a number of sets of auxiliary variables beforehand (including interactions) and to add all variables to the models. One should restrict to demographic and socio-economic characteristics that are generally available in many surveys. When comparing a survey in time, we recommend to fix a number of sets of auxiliary variables. However, now the sets may also include variables that correlate to the main survey items, and variables that relate to the data collection (paradata). When many variables are available, parsimonious models may be favoured.

Partial R-indicators provide insight that is helpful in the reduction of nonresponse. We provide the following simple guidelines:

- In the comparison of different surveys, partial R-indicators are supplementary to R-indicators. Response models are simple and employ general auxiliary variables only.
- In the comparison of a survey in time, partial R-indicators are again supplementary to R-indicators. Response models may be more complex, e.g. define multiple model equations or levels, and may employ paradata additionally to auxiliary variables.
- Conditional partial R-indicators should be used in conjunction with unconditional partial R-indicators. They are always smaller than the unconditional partial R-indicators and comparing the two shows to what extent the apparent impact of a single variable is taken away by the others.
- When many variables are added to models for response, then conditional partial R-indicators naturally are smaller. When two or more variables are included that correlate strongly, then the conditional partial R-indicators will be small for both variables. It is recommendable not to include many related variables.

As a general guideline we conclude with the remark that in improving representativity of response it must always be the objective to increase the response rate and to decrease the R-indicators simultaneously.

12. Visualising R-indicators in R-cockpit

Partial R-indicators are easier to interpret when they are visualised. The R-cockpit program developed in the project RISQ is a graphical tool that enables a quick and easy display of both unconditional and conditional R-indicators. R-cockpit is available at the RISQ website www.risq-project.eu. It is written in R and assumes that the survey data set is converted to R. With the program an R function called *export.R* is provided that executes export of SPSS and SAS data files to R. We refer to the R-cockpit manual for further details.

13. Future releases of `RISQ_R-indicators` in SAS and R

Future releases of RISQ_R-indicators are planned. In 2014 a third release will be provided on www.risq-project.eu that includes population-based R-indicators. Population-based R-indicators measure representativeness based on population counts and population tables only. They widen the scope of the indicators to settings where samples cannot be linked to administrative data. Population-based R-indicators are discussed in

- Shlomo, N., Skinner, C., Schouten, B., Heij, V. de, Bethlehem, J., Ouwehand, P. (2009), Indicators for representative response based on population totals, RISQ deliverable 2.2