sdcLog

Toolkit for Output Control in Research Data Centres

Matthias Gomolka

Deutsche Bundesbank, Research Data and Service Centre (RDSC)

Who am I?

And why do I talk about sdcLog?

I work in the Bundesbank's Research Data and Service Centre.

What I do:

- Production of research data sets on securities transactions
- Data Production Pipelines
- **R tools, which make the RDSC life easier**

Disclaimer:

- No expert in Output Control. I just implemented functionality which we already have for Stata in R.

Motivation

Problem

- Researchers need to show that their output complies to our rules.
- That get's complicated quickly.
- It would be very time-consuming for the RDSC if we would have to check *how* a researcher proved that her output complies to RDSC rules.

Solution

- RDSC provides tools which help researchers to show that their output complies to the rules: **sdcLog**

Theory

Two simple rules:

1. Each result must be based on at least 5 distinct entities (distinct ID's).
2. The two largest entities must not account for more than 85% of a result (n,k-dominance).

Example

A researcher wants to publish the mean of a variable grouped by `sector`. To do so, she has to use `sdc_descriptives()` to show that the output complies to RDSC rules.

```
head(DT)
##     id sector year      val_1    val_2
## 1:   A     S1 2019         NA 9.477642
## 2:   A     S1 2020 94.174449 5.856641
## 3:   B     S1 2019  4.349115 3.697140
## 4:   B     S1 2020  2.589011 6.796527
## 5:   C     S1 2019  6.155680 7.213390
## 6:   C     S1 2020  7.183206 5.948330
```

```
# result
DT[, .(mean = mean(val_1, na.rm = TRUE)),
   by = "sector"]
##     sector     mean
## 1:      S1 15.42511
## 2:      S2 24.43726
```

```
# Proof, that the result complies to rules
sdc_descriptives(DT, id_var = "id", val_var = "val_1", by = "sector")
## OPTIONS: sdc.n_ids: 3 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85
## SETTINGS: id_var: id | val_var: val_1 | by: sector | zero_as_NA: FALSE
## Output complies to RDC rules.
```

Another example

This time, researches want to calculate the result grouped by `sector` and `year`.

```r
sdc_descriptives(DT, id_var = "id", val_var = "val_1", by = c("sector", "year"))
## Warning: DISCLOSURE PROBLEM: Dominant entities.
## OPTIONS: sdc.n_ids: 3 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85
## SETTINGS: id_var: id | val_var: val_1 | by: c("sector", "year") | zero_as_NA: FALSE
## Dominant entities:
##    sector year value_share
## 1:     S2 2020   0.9056314
## 2:     S1 2020   0.8776852
## 3:     S1 2019   0.6815011
## 4:     S2 2019   0.5506965
```

Minimum and maximum values

Now, researchers want to publish minimum and maximum values as well.

Problem

Minimum and maximum value are confidential micro data.

Solution

"Minimum" and "maximum" value as mean of `n` smallest / largest values using `sdc_min_max()`:

```
sdc_min_max(DT, id_var = "id", val_var = "val_1")
## OPTIONS: sdc.n_ids: 3 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85
## SETTINGS: id_var: id | val_var: val_1
##    val_var      min distinct_ids_min      max distinct_ids_max
## 1:   val_1 2.320075                3 37.34043                7
```

Output control for models

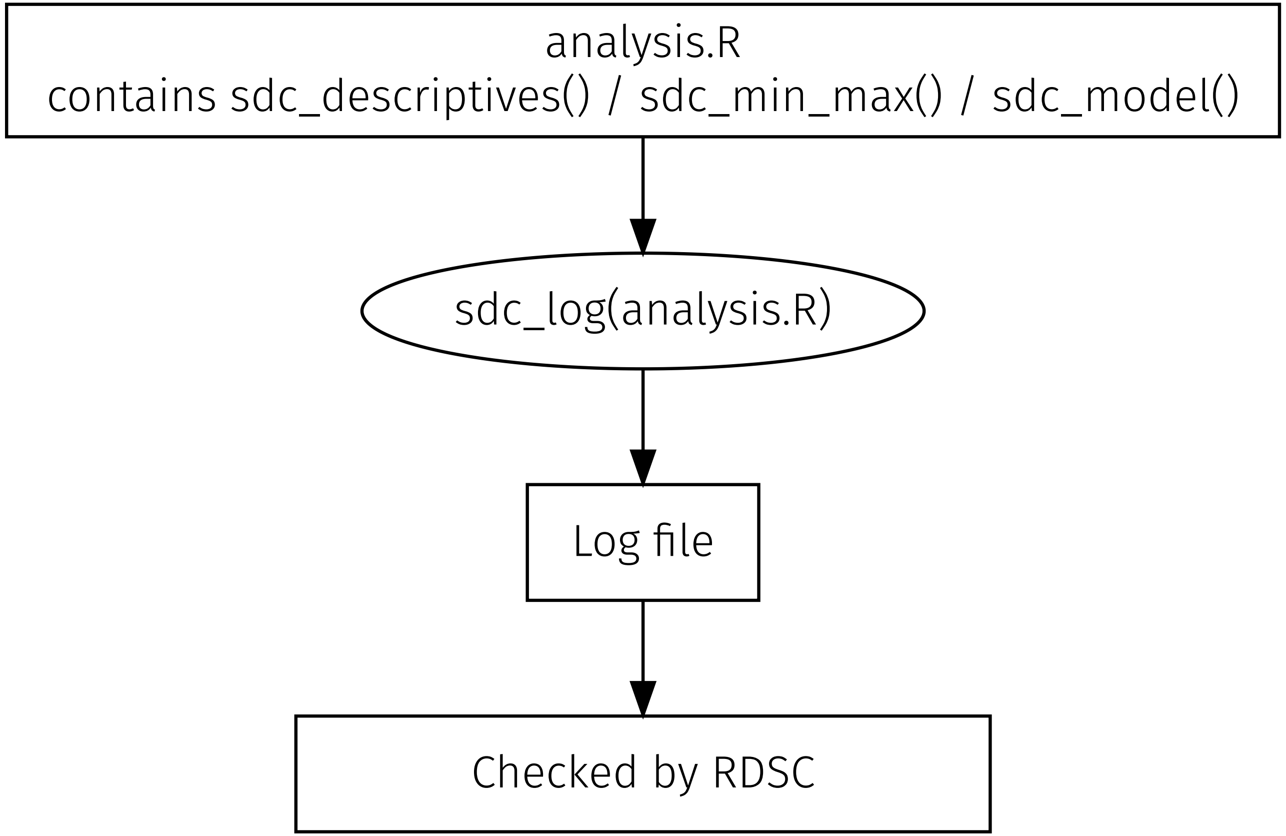
Researchers also want to publish results from a linear regression.

```
options(sdc.n_ids = 3)

# Estimate model
mod <- lm(val_1 ~ sector + year + val_2, data = DT)

# Check if model complies to rules
sdc_model(DT, model = mod, id_var = "id")
## OPTIONS: sdc.n_ids: 3 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85
## SETTINGS: id_var: id
## Output complies to RDC rules.
```

Why is it called sdcLog?



analysis.R
contains sdc_descriptives() / sdc_min_max() / sdc_model()

sdc_log(analysis.R)

Log file

Checked by RDSC

Installation und contact information

CRAN

```r
install.packages("sdcLog")
```

GitHub

https://github.com/matthiasgomolka/sdcLog/issues

E-mail

matthias.gomolka@bundesbank.de

Phone

069 9556-4991