sdcLog

Werkzeuge für Outputkontrolle in Forschungsdatenzentrem

Matthias Gomolka

Deutsche Bundesbank, Forschungsdaten- und Servicezentrum

Wer bin ich?

Ich arbeite im Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank.

Inhaltliche Schwerpunkte:

- Wertpapiertransaktionsdaten
- Data Production Pipelines
- **R-Tools, die das FDSZ-Leben einfacher machen**

Motivation

Problem

- Forschende stehen in der Pflicht, zu zeigen, dass ihr Output den Regeln des FDSZ entspricht.
- Das kann schnell sehr komplex werden.
- Außerdem ist der Aufwand für das FDSZ sehr hoch, wenn zusätzlich geprüft werden muss, wie die Forschenden die Nachweise für Ihren Output erbringen.

Lösung

- FDSZ stellt Forschenden Werkzeuge zur Verfügung, um die Konformität mit den Outputregeln nachzuweisen: **sdcLog**

Theorie

Zwei einfache Regeln:

1. Jedes Ergebnis muss auf mindestens 5 unterschiedlichen Entitäten basieren (distinct ID's).
2. Die beiden größten Entitäten dürfen nicht mehr als 85% eines Ergebnisses ausmachen (dominance).

Ein Beispiel

Forschende möchten das arithmetische Mittel einer Variable berechnen und das Ergebnis in ihrer Publikation zeigen. Dazu müssen sie vorab mit `sdc_descriptives()` zeigen, dass das Ergebnis den Output-Regeln entspricht.

```
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
```

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

```
# Nachweis, dass das Ergebnis den Regeln entspricht
sdc_descriptives(DT, id_var = "id", val_var = "val_1", by = "sector")
## OPTIONS: sdc.n_ids: 5 | 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.
```

Präsentation verfügbar unter https://tinyurl.com/sdcLog-FDI2021

Noch ein Beispiel

Diesmal berechnen die Forschenden das arithmetische Mittel gruppiert nach `sector` und `year`.

```
sdc_descriptives(DT, id_var = "id", val_var = "val_1", by = c("sector", "year"))
## Warning: DISCLOSURE PROBLEM: Not enough distinct entities.
## Warning: DISCLOSURE PROBLEM: Dominant entities.
## OPTIONS: sdc.n_ids: 5 | 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
## Not enough distinct entities:
##    sector year distinct_ids
## 1:     S1 2019            4
## 2:     S1 2020            5
## 3:     S2 2019            5
## 4:     S2 2020            5
## 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 und Maximum

Jetzt möchten Forschende neben dem arithmetischen Mittel auch noch das Minimum und Maximum einer Variablen zeigen.

Problem

Minimum und Maximum sind vertrauliche Einzeldaten.

Lösung

"Minumum" und "Maximum" als arithmetisches Mittel der kleinsten bzw. größten Werte mit `sdc_min_max()`:

```
sdc_min_max(DT, id_var = "id", val_var = "val_1")
## OPTIONS: sdc.n_ids: 5 | 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 3.364432                5 37.34043                7
```

Outputkontrolle bei statistischen Modellen

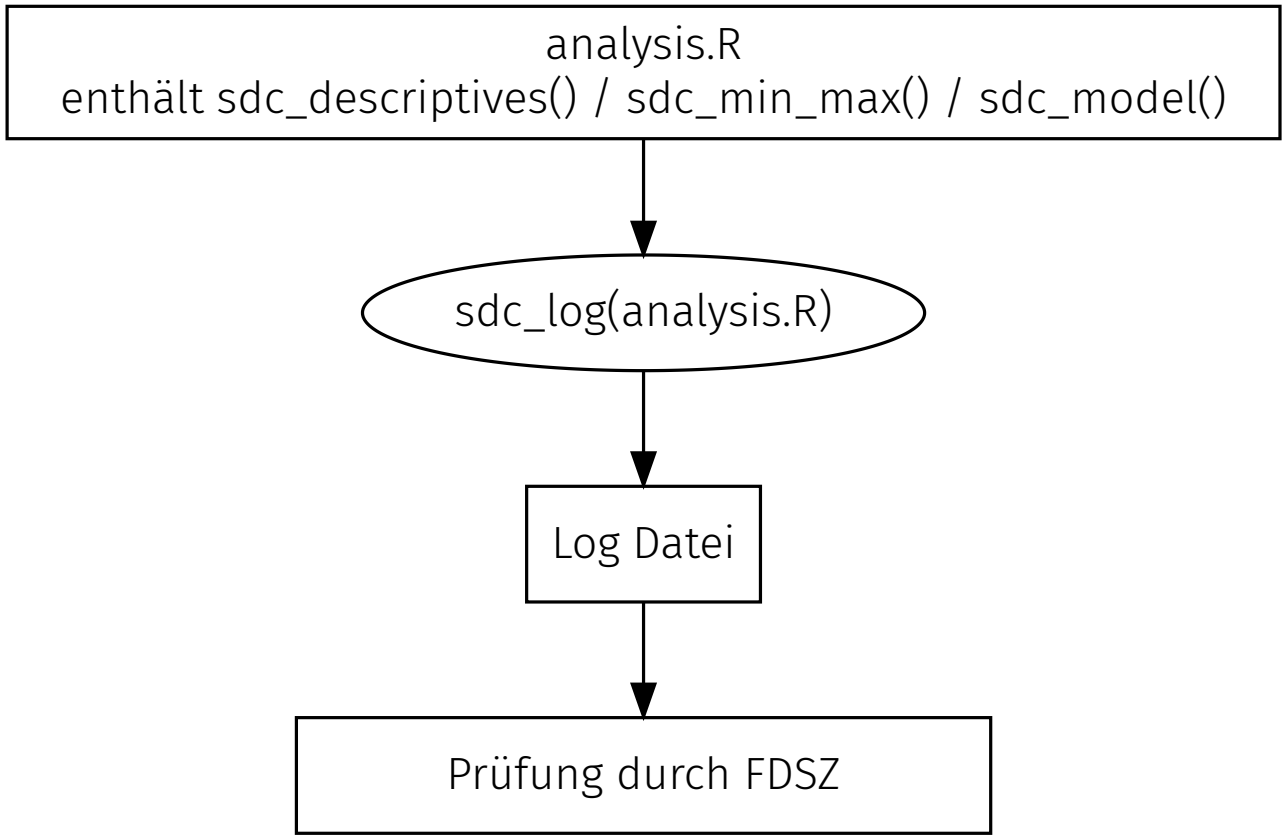
Jetzt möchten Forschende die Ergebnisse einer linearen Regression veröffentlichen.

```
options(sdc.n_ids = 3)

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

# Prüfen, ob Ergebnise freigegeben werden können
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.
```

Warum heißt es sdcLog?



Installation und Kontakt

CRAN

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

GitHub

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

E-Mail

matthias.gomolka@bundesbank.de

Telefon

069 9556-4991

Präsentation verfügbar unter https://tinyurl.com/sdcLog-FDI2021