An R implementation of a model-based estimator – a UK study

Konstantinos Soulanis
Traditional approaches to survey inference

- Design-based approach (Horvitz-Thompson estimator)
- Model-assisted approach (ratio estimator)
- Well-understood properties
- Available in *survey* and *ReGenesees* libraries
Annual Survey of Goods and Services (ASGS)

• Sample size ~40,000 UK businesses
• Over 2000 service products (study variables)
• Estimates for each product produced for each service industry class (4-digit level of SIC)
• Outputs to be used for important economic indicators such as GDP
Model-based approach

- Two-part conditional ratio, also called Chambers-Cruddas model (CC)
- Part 1: estimate the probability that a business provides a particular service
- Part 2: model the service turnover as proportional to register employment

\[
\hat{t}_y = \sum_{i \in s} y_i + \sum_{i \notin s} y_i
\]
Using R: custom functions

- Successive functions estimate the model parameters, and then the totals, with variances and standard errors
- Outputs of certain functions are used as inputs to other ones
- Tracking of calculations, debugging, convenient in outliers treatment
Pipeline flow chart

- Data preparation
- Propensities calculation
- Variance and standard errors calculations
- Estimates calculation
- Final output

• Calculation of the slopes of the fitted lines
• Other parameters for the variance
• Trimming is applied for outliers treatment
Code Examples – Data preparation

```r
finaldata <- finaldata %>%
  left_join(popcounts, by = "cell_no") %>%
  select(RUReference:sizeband, ncount, bign, everything()) %>%
  mutate_at(vars(RUReference, FormType, cell_no, sic2007, sic_group),
            funs(as.numeric(.))) %>%
  mutate(sic4 = as.numeric(substr(sic2007, 1, 4)))
  select(RUReference:sic2007, sic4, everything())

finaldata_prep <- finaldata %>%
  mutate_at(vars(-RUReference:-bign),
            funs("DELTA" = ifelse(.>0, 1, 0),
                            "YxD_X" = ifelse(Emptfro > 0, .^2/Emptfro, 0))) %>%
  mutate_at(vars(contains("DELTA")), funs("X"=.^Emptfro))

responses_sic4 <- finaldata_mod %>%
  group_by(sic4) %>%
  summarise_at(vars(starts_with("UK"), starts_with("OS")),
               funs(sum(.!=0))) %>%
  gather(question, responses, -sic4) %>%
  filter(responses!=0)
```
Code Example - custom functions

calc_betas <- function(df, grp.var){

  grp.var <- enquo(grp.var)

  ## Betas formula: \( \beta = \frac{\sum(D \cdot Y)}{\sum(D \cdot X)} \):

  # numerator:
  betas.num <- df %>%
    group_by(cell_no, !!grp.var) %>%
    summarise_at(vars(-RUReference,:bign, -contains("DELTA"), -contains("YxD_X")),sum) %>%
    group_by(!!grp.var) %>%
    summarise_at(vars(-cell_no),sum) %>%
    arrange(!!grp.var)

  # denominator:
  betas.denom <- df %>%
    group_by(cell_no, !!grp.var) %>%
    summarise_at(vars(contains("DELTA_X")),sum) %>%
    group_by(!!grp.var) %>%
    summarise_at(vars(-cell_no),sum) %>%
    arrange(!!grp.var)

  # betas final:
  BETAS <- cbind(betas.num[1], betas.num[-1] / betas.denom[-1])

  return(BETAS)
}
Scatterplots of Standard Errors

At least 2 responses
Scatterplots of Standard Errors

At least 2 responses

At least 5 responses

At least 10 responses

At least 15 responses

At least 20 responses

At least 25 responses

At least 30 responses

At least 50 responses

At least 70 responses

Expansion Standard Errors (x1,000)

CC-model Standard errors (x1,000)
Ratio of Standard Errors

Log10-Ratio (CC-model / Exp)

number of non-zero responses per industry-service combination
Conclusions / Future work

- CC estimator is more efficient than traditional expansion estimator
- R offered the versatility and speed to handle large amount of data and calculations
- Several functions and procedures were created, contributing to the Office’s transition towards open software
- The code will continue to be tested, improved and finally packaged to be made generally available
- Further work for improved sampling and sample allocation
Thank you!

Contact details: konstantinos.soulanis@ons.gov.uk