Office for National Statistics

Evaluation of estimation methods for a new survey of the UK's Office for National Statistics using R

Konstantinos Soulanis

Outline

- Background of the survey
- Estimation methods
- Model implementation using R
- Results
- Conclusions
- Recommendations and further work

Annual Survey of Goods and Services (ASGS)

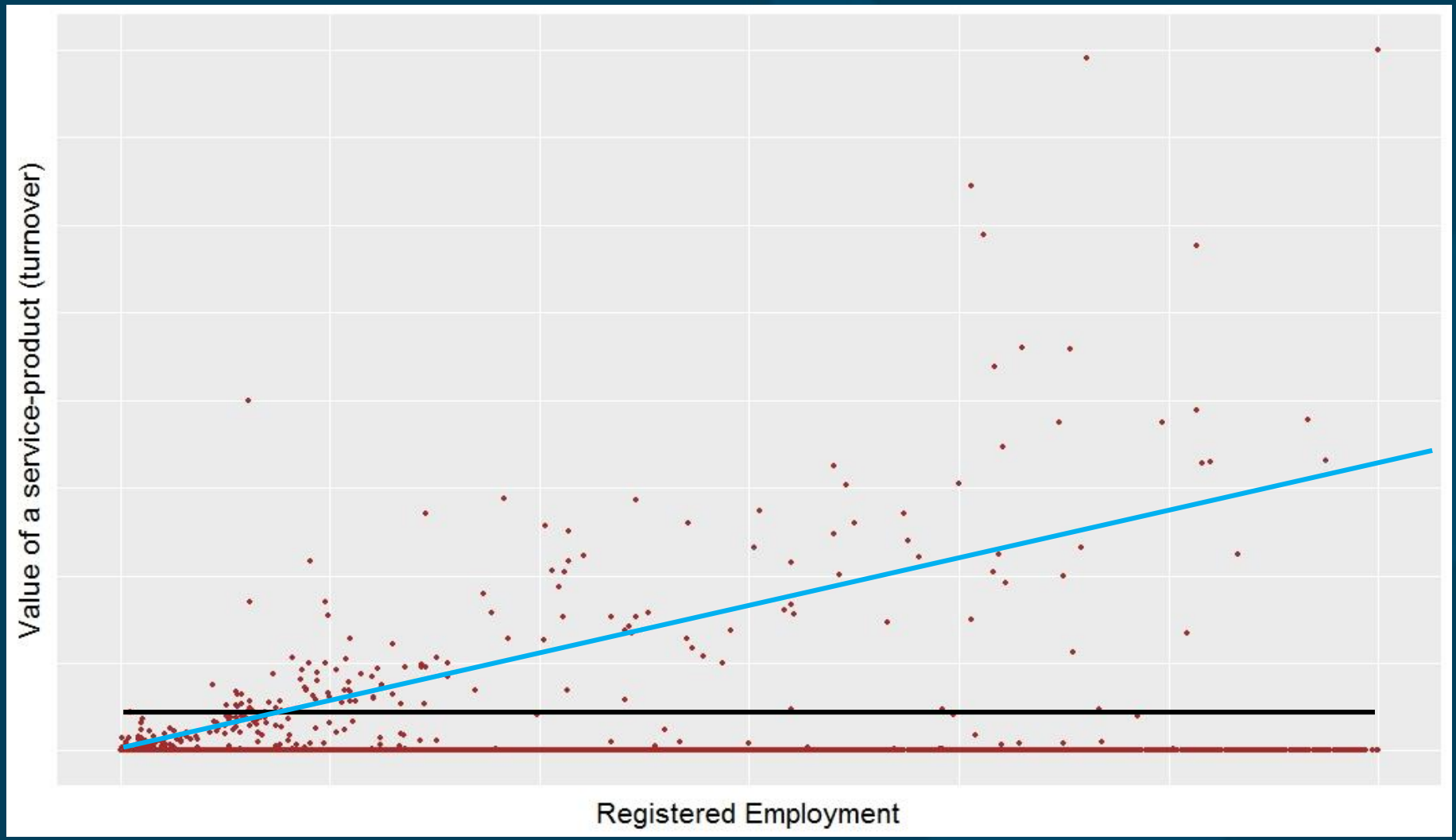
- Sample size ~40,000 UK businesses
- Questions on >2000 service products, divided into domestic and export markets
- Estimates for each product produced for each service industry class (4-digit level of SIC)
- Outputs to be used for important economic indicators such as GDP

Estimation method 1: Expansion

- Simple estimator:

$$\hat{Y} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in h}^{n_h} y_i$$

- Well-understood properties
- Available in ReGenesees, but simple to code manually
- Used in previous similar survey


Value of a service-product (turnover)
Registered Employment

Method 2: model-based conditional ratio

- 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

Model fitting

- Parameter estimates:

$$\hat{\pi}_{cah} = \frac{1}{n_{ah}} \sum_{i \in s_{ah}} \Delta_{cahi} \qquad \hat{\beta}_{ca} = \frac{\sum_h \sum_{i \in s_{ah}} \Delta_{cahi} Y_{cahi} X_{ahi} / \upsilon_{cah}(X_{ahi})}{\sum_h \sum_{i \in s_{ah}} \Delta_{cahi} X_{ahi}^2 / \upsilon_{cah}(X_{ahi})}$$

$$\hat{\sigma}_{ca}^2 = \frac{1}{\sum_h \sum_{i \in s_{ah}} \Delta_{cahi} - 1} \sum_h \sum_{i \in s_{ah}} \frac{\Delta_{cahi} (Y_{cahi} - \hat{\beta}_{ca} X_{ahi})^2}{\upsilon_{cah}(X_{ahi})}$$

- Predictor for the total production of product c by industry a:

$$\hat{T}_y(a,c) = \sum_h \sum_{i \in s_{ah}} Y_{cahi} + \hat{\beta}_{ca} \sum_h \hat{\pi}_{cah} \sum_{i \in r_{ah}} X_{ahi}$$

- Total predicted production of product-service *c:*

$$\hat{T}_y(c) = \sum_a \hat{T}_y(a,c)$$

Variances: Expansion vs CC-model

- Expansion:

$$Var_{(\exp)} = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} s_h^2 \qquad \text{where,} \qquad s_h^2 = \sum_{i \in h}^{n_h} \frac{(y_i - \bar{y})^2}{n_h - 1}$$

- CC-model:

$$Var\{\hat{T}(a,c) - T(a,c) \mid X\} \approx$$

$$\approx \beta_{ca}^2 \sum_h \frac{\pi_{cah}(1-\pi_{cah})}{n_{ah}} \left(\sum_{i \in r_{ah}} X_{ahi}\right)^2 + \beta_{ca}^2 \sum_h \pi_{cah}\{1-\pi_{cah}\} \sum_{i \in r_{ah}} X_{ahi}^2 +$$

$$+ \frac{\sigma_{ca}^2}{\sum_h \sum_{i \in s_{ah}} \pi_{cah} X_{ahi}^2 / \upsilon_{cah}(X_{ahi})} \left\{ \sum_h \frac{\pi_{cah}(1-\pi_{cah})}{n_{ah}} \left(\sum_{i \in r_{ah}} X_{ahi}\right)^2 + \left(\sum_h \pi_{cah} \sum_{i \in r_{ah}} X_{ahi}\right)^2 \right\} +$$

$$+ \sigma_{ca}^2 \sum_h \pi_{cah} \sum_{i \in r_{ah}} \upsilon_{cah}(X_{ahi})$$

Using R: Data preparation

- Main dataset consists of 24,968 observations of 2,078 are the service-products
- Universe file consists of 1,878,710 businesses
- Using libraries dplyr and tidyr we prepare the datasets, in order to be used in the custom functions

Code Examples

```
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)
```

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

Code Example

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

  grp.var <- enquo(grp.var)

  ## Betas formula: [ b = Sumof(D*Y) / Sumof(D*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)

}
```

Using R: Outlier treatment

CC-model:

- Trimming of the 5% top/bottom extreme non-zero values
- Propensities remain the same
- A custom function to perform trimming by industry and product
- Subsequent calculated through a loop

Expansion estimation:

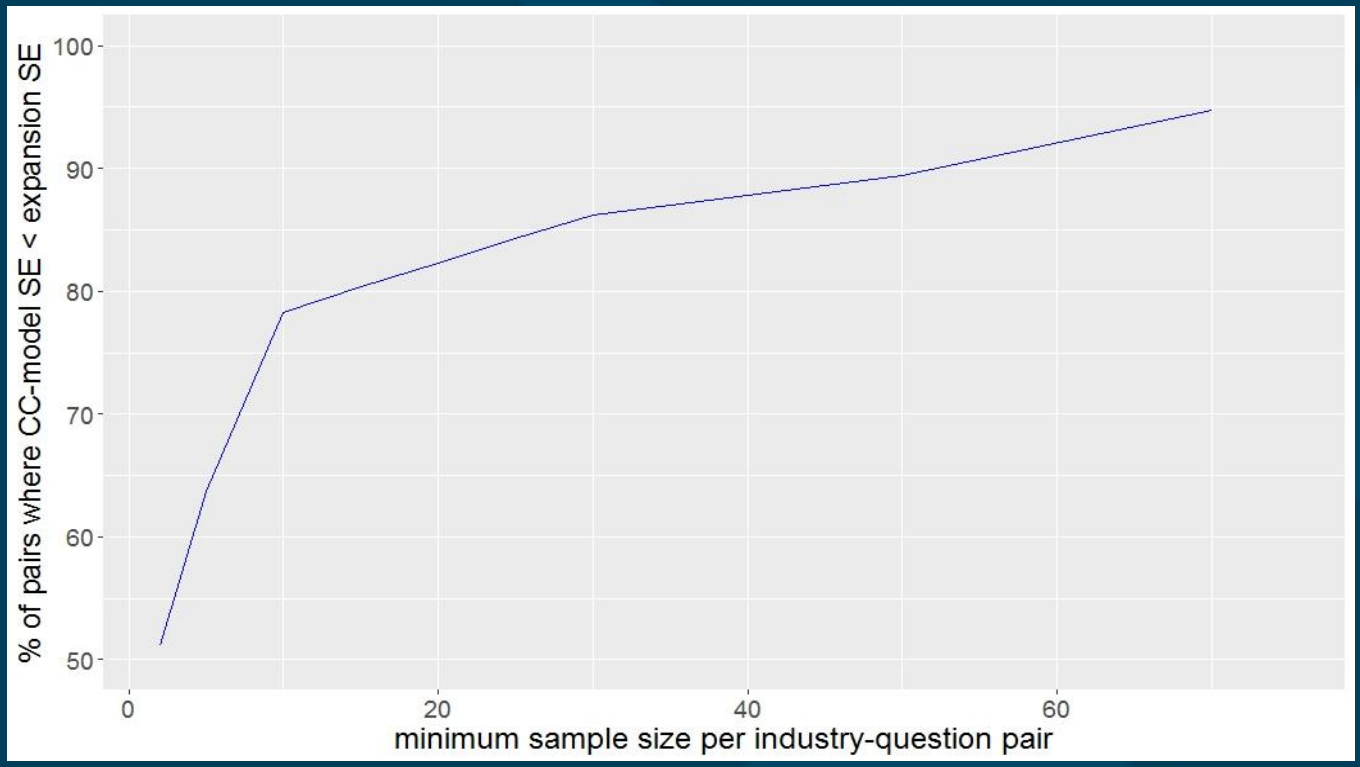
- Winsorisation method
- Reduces the impact of outliers while minimising MSE
- Method coded in R, creating two custom functions and two loops

Results

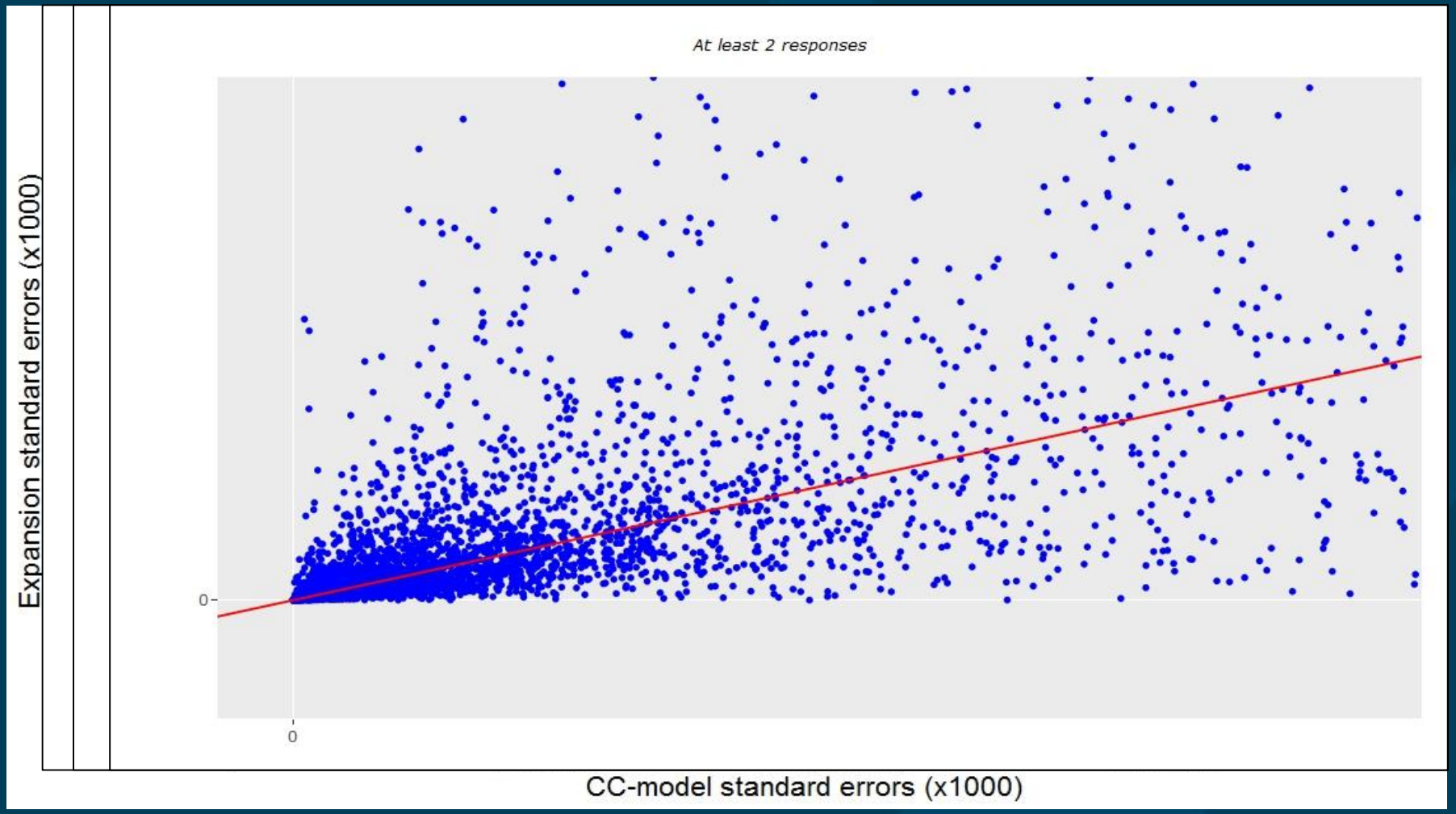
- Evaluate methods by comparing standard errors
- Libraries ggplot2, gridExtra and plotly for visualisation
- Custom functions for quick filtering/plotting of data
- Allow filtering by number of non-zero responses

Results

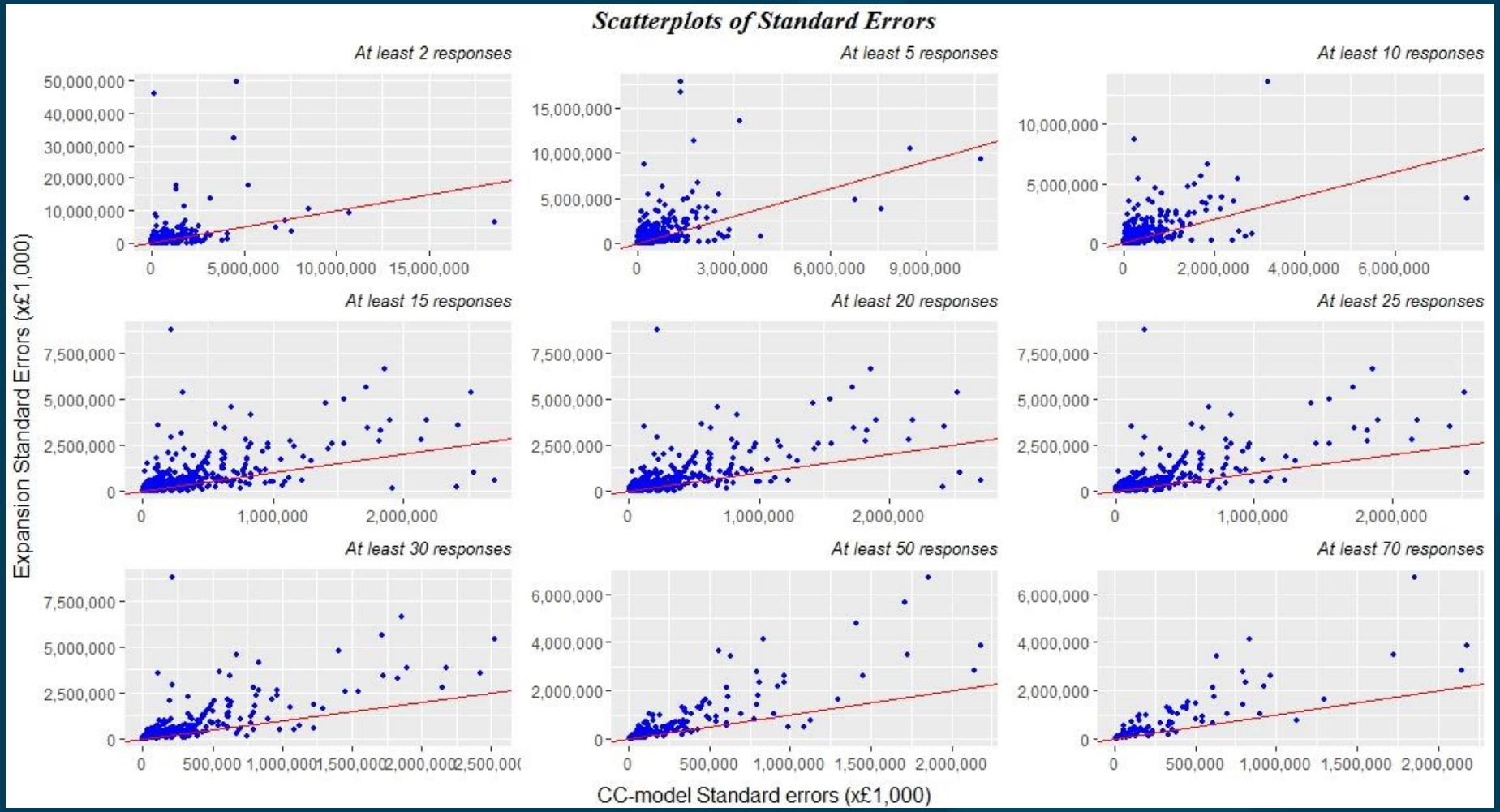
Minimum number of non-zero responses	% of pairs where CC has smaller S.E. than expansion
2	51.1
5	63.7
10	78.3
15	80.4
20	82.3
30	86.2
50	89.4
70	94.8



Results – Scatterplots of Standard Errors



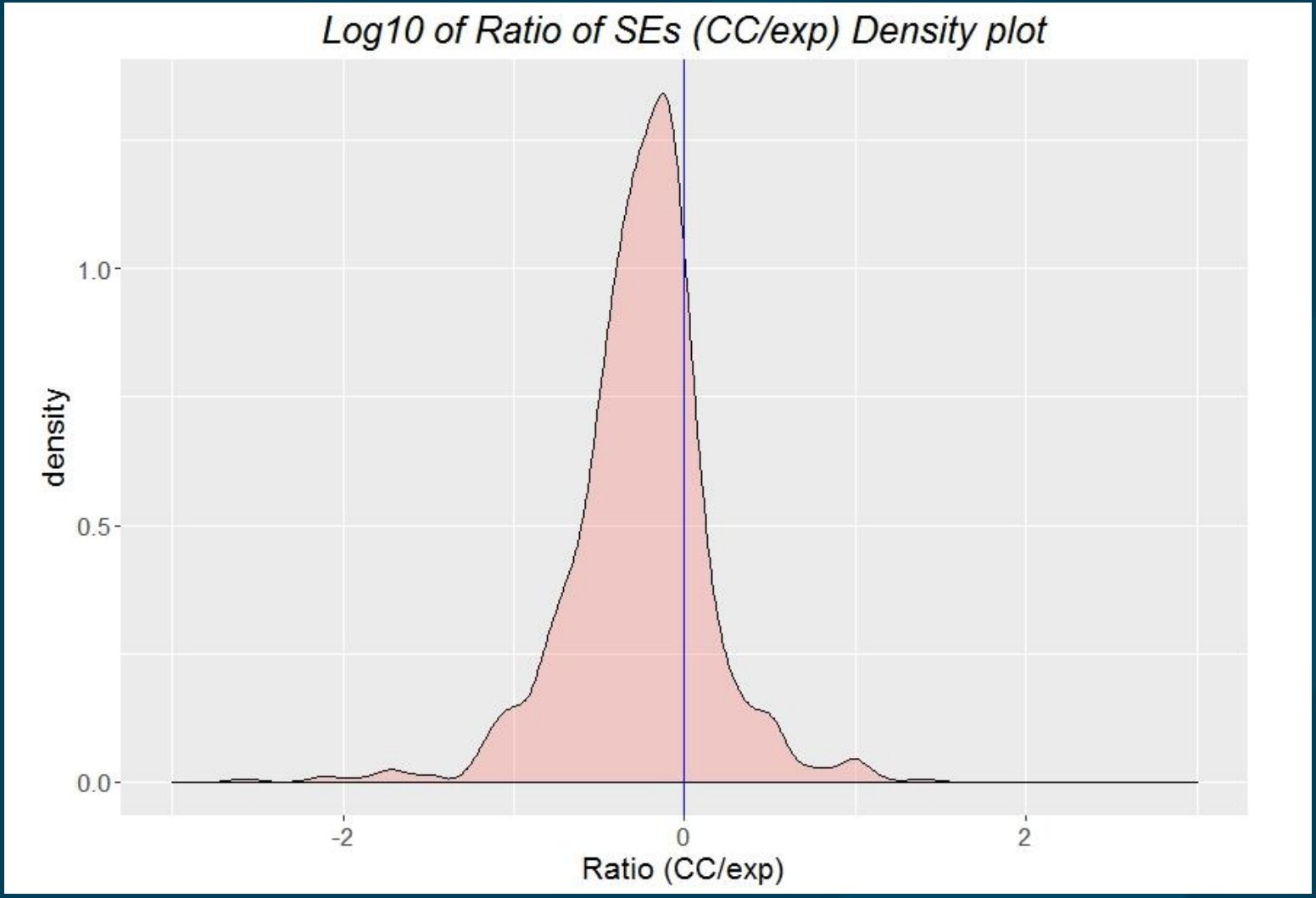
Results – Scatterplots of Standard Errors



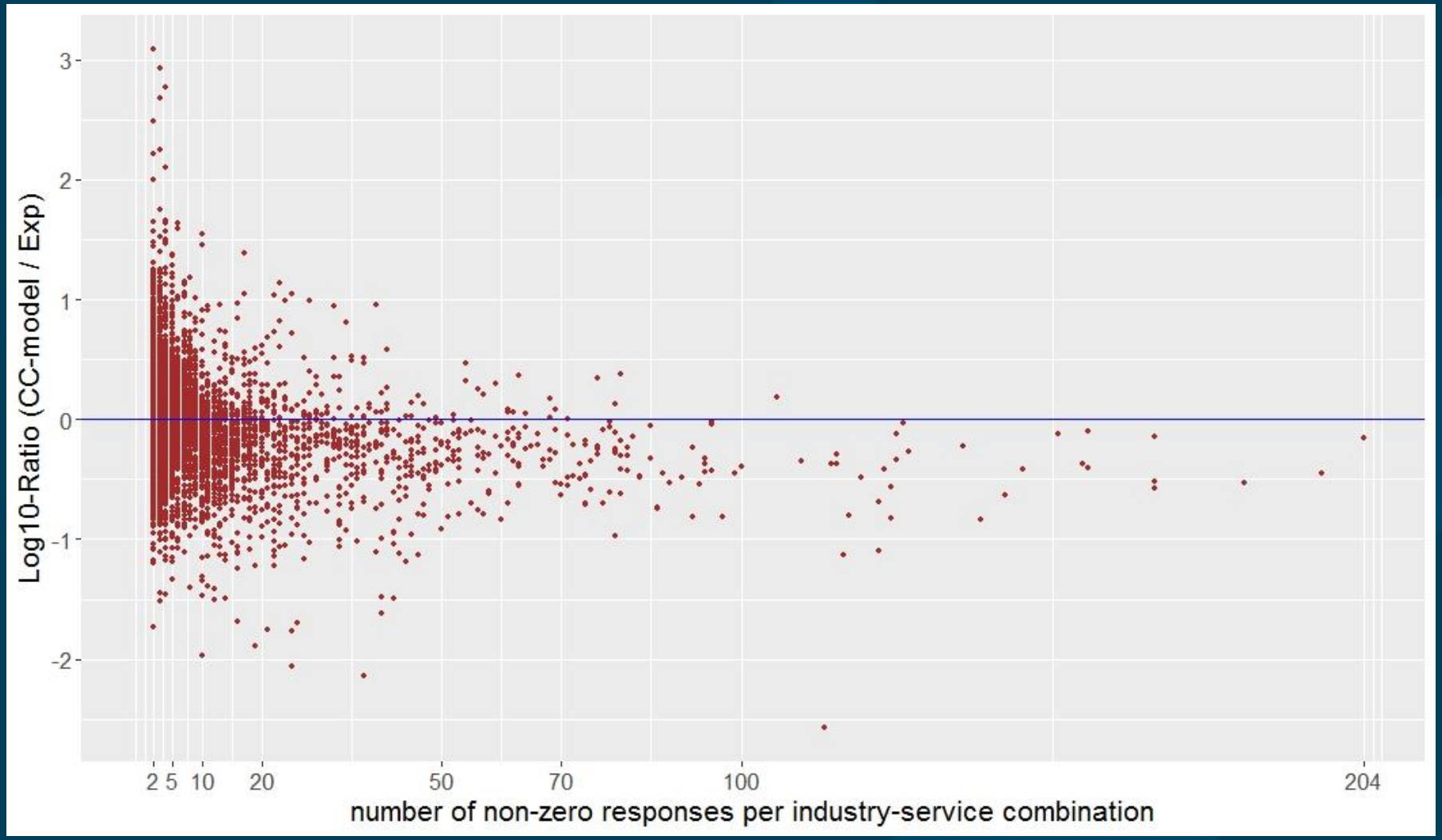
Results – Ratio of Standard Errors

- We can observe the difference between the methods with more precision
- Can be quantified using the logarithm of the ratio
- Formula used: $$Ratio_{ca} = \frac{S.E._{ca}^{\text{CC-model}}}{S.E._{ca}^{\text{expansion}}}$$
- If this value is smaller than 1, then CC-model is preferable to expansion

Results – Ratio of Standard Errors



Results – Ratio of Standard Errors



Conclusions

- As the number of non-zero responses increases, CC estimator becomes more efficient than expansion
- Treatment of outliers improved the overall results and useful methods were developed in R
- 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

Recommendations and further work

- Survey team was recommended to use the CC-model for the production of the ASGS 2016 estimates (first release: 31/8/2018)
- The code will continue to be tested, improved and finally packaged to be made generally available
- Further work for improved sampling and sample allocation

References and useful links

- Chambers R. and Clark G. (2012). *An introduction to Model-Based Survey Sampling with Applications*. Oxford statistical science series No.37
- ASGS latest publication: https://www.ons.gov.uk/businessindustryandtrade/business/businesservices/bulletins/annualsurveyofgoodsandservices/2016
- ASGS development: https://www.ons.gov.uk/businessindustryandtrade/business/businesservices/articles/developmentoftheannualsurveyofgoodsandservices/2018-08-31
- Various ONS's internal reports
- The enormous R community on the internet!

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

Thank you for your time!

QUESTIONS?