

README Example

Generalized Expectile Regression with Flexible Response Function

Elmar Spiegel & Thomas Kneib & Petra von Gablenz & Fabian Otto-Sobotka

06 12 2020

Contents

1	General	1
2	Analysis of self-reported hearing	1
2.1	Content	1
2.2	For checking reproducibility	2
2.3	Session Info	2
2.4	Variable descriptions	2
3	Modelling of mercury concentration in blood	3
3.1	Content	3
3.2	For checking reproducibility	4
3.3	Session Info	4
3.4	Variable descriptions	5

1 General

- This README describes the examples of the manuscript.
- Each example has it's own directory and it's own chapter in this README.

2 Analysis of self-reported hearing

- Files are stored in directory `Code_Example_Hearing`

2.1 Content

- The data set used in the manuscript is confidential, however a fake data set can be produced which mimics the original behavior.
- The data result from the HÖRSTAT study. For details see *von Gablenz and Holumbe (2017)*
- We provide two kinds of analysis, one where the smoothing parameters fixed at the 50% asymmetry level, one where the smoothing parameters are optimized for each asymmetry level separately.
- .R-Files:
 - `Code_Example_Hearing_50.R`: File to estimate those models, where the smoothing parameters are fixed on the 50% asymmetry level. Only thing necessary for the figures in the main paper. Generates `Figure4.eps` and `Figure5.eps`, along with results displayed in the supplementary materials, e.g. the Sections C.1.1 and C.1.2 and C.2.

- `Code_Example_Hearing_all.R`: File to estimate those models, where the smoothing parameters are optimized for each asymmetry level separately. All results are displayed only in the supplementary materials, e.g. the Sections C.1.3 and C.1.4 and C.2.
 - `BuildExampleData.R`: File to produce the fake data. Used for checking reproducibility.
 - `Legends1.R`: File to produce the legends used in the supplementary materials.
- Please be aware that you need to open the files with Encoding UTF-8, due to some German Umlaute in the data set / code.

2.2 For checking reproducibility

1. Run `BuildExampleData.R` to produce the fake data.
2. Change in `Code_Example_Hearing_50.R` from the original to the fake data set as indicated in line 17-27.
3. Run `Code_Example_Hearing_50.R` until the start of the bootstrap line 406 (duration approx. 1.5h) The bootstrap takes longer, especially the part from line 465 onwards (duration approx. 10h)

2.3 Session Info

```
> sessionInfo()
R version 3.6.1 (2019-07-05)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Mojave 10.14.6

Matrix products: default
BLAS: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
[1] de_DE.UTF-8/de_DE.UTF-8/de_DE.UTF-8/C/de_DE.UTF-8/de_DE.UTF-8

attached base packages:
[1] stats graphics grDevices utils datasets methods base

other attached packages:
[1] colorspace_1.4-1 FlexER_0.04-06 mgcv_1.8-29 nlme_3.1-140 Matrix_1.2-17
[6] MASS_7.3-51.4 foreign_0.8-72

loaded via a namespace (and not attached):
[1] compiler_3.6.1 RcppEigen_0.3.3.5.0 tools_3.6.1 Rcpp_1.0.2
[5] splines_3.6.1 grid_3.6.1 lattice_0.20-38
```

2.4 Variable descriptions

- **SSQ_ListeningEffort: Response variable** Self-reported hearing abilities as collected in the Speech, Spatial and Qualities of Hearing Scale (SSQ Gatehouse and Noble, 2004) which was used in its German short form. In this paper, we concentrate on one single item which raises the question of listening effort: “Do you have to put in a lot of effort to hear what is being said in conversation with others?”. Ranging from 0 to 10.
- **Alter:** Age of the persons in years. Ranging from 18 to 97
- **GOESA_BE:** In examining hearing performance, the participants conducted the Göttingen Sentence Test in noise (GÖSA, Kollmeier and Wesselkamp, 1997). Everyday sentences were presented in background noise to the participants whose task was to repeat as many sentences or single words as possible. The speech level was continuously adapted to the individual participants performance to estimate the speech reception threshold (SRT) in decibel signal-to-noise ratio which refers to a speech intelligibility of 50%. GOESA_BE represents the better-ear SRT for each individual. Ranging from -7.8 to 4.

- `GOESA_diff`: The absolute difference between the left and the right ear SRT as a quantification of hearing asymmetry in decibel. Ranging from 0 to 6.4
- `Schule_agg_SSQ`: Educational level. 1 represents a basic German education level, 2 represents intermediate education, 3 represents a high level of general education
- `Gender`: Gender of the person. 0 represents “female”, 1 represents “male”.

3 Modelling of mercury concentration in blood

- Files are stored in directory `Code_Example_THg`

3.1 Content

- The data are downloaded from <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx> at 04.11.2020 more details see below.
- We provide two kinds of analysis, one where the smoothing parameters fixed at the 50% asymmetry level, one where the smoothing parameters are optimized for each asymmetry level separately.
- `.R-Files`:
 - `Code_Example_THg_data_prep.R`: Data preparation:
 - * Loading of the XPT data
 - * Merging the different data files by `seqn` variable
 - * Building of new variables
 - * Removing of missing values
 - * Output: Final data set `Example_THg.RData`
 - `Code_Example_THg_model_50.R`:
 - * File to estimate those models, where the smoothing parameters are fixed on the 50% asymmetry level.
 - * Only thing necessary for the figure in the main paper.
 - * Generates `Figure6.eps`, along with results displayed in the supplementary materials, e.g. the Sections D.1.1 and D.1.2 and D.2.
 - `Code_Example_THg_model_all.R`:
 - * File to estimate those models, where the smoothing parameters are optimized for each asymmetry level separately.
 - * All results are displayed only in the supplementary materials, e.g. the Sections D.1.3 and D.1.4 and D.2.
- `.XPT-Files`: Original data
 - `BMX_G.XPT`
 - * NHANES body measures data
 - * Source: https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/BMX_G.XPT
 - * Description: https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/BMX_G.htm
 - * Variables used:
 - `bmxbmi` \Rightarrow BMI in $\frac{kg}{m^2}$
 - `DEMO_G.XPT`
 - * NHANES demographic variables
 - * Source: https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/DEMO_G.XPT

- * Description: https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/DEMO_G.htm
 - * Variables used:
 - `riagendr` ⇒ Gender
 - `ridageyr` ⇒ Age in years
 - `dmdeduc2` ⇒ Education level
 - `indhhin2` ⇒ Household Income in categories
- DR1TOT_G.XPT
- * NHANES Dietary Interview
 - * Source: https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/DR1TOT_G.XPT
 - * Description: https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/DR1TOT_G.htm
 - * Variables used:
 - `drd350aq`, `drd350bq`, ..., `drd350jq` ⇒ Frequency of eating different kind of shellfish in the past 30 days
 - `drd370aq`, `drd370bq`, ..., `drd370uq` ⇒ Frequency of eating different kind of fish in the past 30 days
- PBCD_G.XPT
- * NHANES Laboratory data on Cadmium, Lead, Total Mercury, Selenium, & Manganese in Blood
 - * Source: https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/PBCD_G.XPT
 - * Description: https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/PBCD_G.htm
 - * Variables used:
 - `lbxthg` ⇒ Total mercury in blood in $\mu\text{g}/\text{L}$ (THg)

3.2 For checking reproducibility

1. Run `Code_Example_THg_data_prep.R` to produce the final data set.
2. Run `Code_Example_THg_model_50.R` until the start of the bootstrap line 348 (duration up to 15min)
The bootstrap takes longer, especially the part from line 419 onwards (duration in total approx. 9h)

3.3 Session Info

3.3.1 Data preparation

```
> sessionInfo()
R version 3.6.3 (2020-02-29)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.1 LTS

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0

locale:
 [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8      LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8     LC_NAME=C
 [9] LC_ADDRESS=C             LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats      graphics  grDevices  utils      datasets  methods   base

other attached packages:
[1] Hmisc_4.4-1  ggplot2_3.3.2  Formula_1.2-4  survival_3.1-12
[5] lattice_0.20-41

loaded via a namespace (and not attached):
 [1] pillar_1.4.6      compiler_3.6.3  RColorBrewer_1.1-2
```

```

[4] tools_3.6.3      base64enc_0.1-3  rpart_4.1-15
[7] digest_0.6.27   lifecycle_0.2.0  tibble_3.0.4
[10] gtable_0.3.0    htmlTable_2.1.0  checkmate_2.0.0
[13] pkgconfig_2.0.3 png_0.1-7        rlang_0.4.8
[16] Matrix_1.2-18   rstudioapi_0.11  xfun_0.19
[19] gridExtra_2.3   stringr_1.4.0    withr_2.3.0
[22] cluster_2.0.9   knitr_1.30       vctrs_0.3.4
[25] htmlwidgets_1.5.2 grid_3.6.3       nnet_7.3-14
[28] data.table_1.13.2 glue_1.4.2       R6_2.5.0
[31] jpeg_0.1-8.1   foreign_0.8-75   latticeExtra_0.6-29
[34] magrittr_1.5    scales_1.1.1     backports_1.2.0
[37] ellipsis_0.3.1  htmltools_0.5.0  splines_3.6.3
[40] colorspace_1.4-1 stringi_1.5.3     munsell_0.5.0
[43] crayon_1.3.4

```

3.3.2 Model estimation

```

> sessionInfo()
R version 3.6.3 (2020-02-29)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.1 LTS

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0

locale:
 [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8      LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8     LC_NAME=C
 [9] LC_ADDRESS=C             LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats      graphics  grDevices  utils      datasets  methods    base

other attached packages:
[1] colorspace_1.4-1 FlexER_0.04-06 mgcv_1.8-31 nlme_3.1-148
[5] Matrix_1.2-18 MASS_7.3-51.6 foreign_0.8-75

loaded via a namespace (and not attached):
[1] compiler_3.6.3      RcppEigen_0.3.3.5.0 Rcpp_1.0.3
[4] splines_3.6.3       grid_3.6.3          lattice_0.20-41

```

3.4 Variable descriptions

- **THg: Response variable** Concentration of mercury in blood (in $\mu g/L$). Ranging from 0.11 to 50.81
- **BMI:** BMI of the person in $\frac{kg}{m^2}$. Ranging from 16.2 to 49.9.
- **Age:** Age of the persons in years. Ranging from 21 to 79.
- **Sex:** Gender of the persons. Categories “male” and “female”. Reference category: male
- **Education:** Educational level of the persons. (What is the highest grade or level of school {you have/SP has} completed or the highest degree {you have/s/he has} received?): Categories College graduate or above (College), Some college or AA degree (AA degree), High school graduate/GED or equivalent (High school), 9-11th grade (Includes 12th grade with no diploma) (9th-11th grade), Less than 9th grade (Less 9th grade). Reference category College.
- **Income:** Total household income in Dollar: Categories 0-20000 (00-20), 20000-75000 (20-75), more than 75000 (75+). Reference category 00-20.
- **Shell_count:** Number of times different kind of shellfish were eaten in the past 30 days. Sum of different kind of shellfish variables. Categories: no shellfish (0), on average up to once a week (1-4), on average up to twice a week (5-8), on average more than twice a (8+). Reference category 0.

- **Fish_count**: Number of times different kind of fish were eaten in the past 30 days. Sum of different kind of fish variables. Categories: no fish (0), on average up to once a week (1-4), on average up to twice a week (5-8), on average more than twice a (8+). Reference category 0.