Основы программирования в R

Выгрузка описательных статистик с помощью `stargazer`

Алла Тамбовцева, НИУ ВШЭ

Библиотека `stargazer` используется для красивой выгрузки результатов из R. Такое интересное название она имеет из-за того, что чаще всего из R экспортируют результаты регрессионных моделей, где статистически значимые коэффициенты отмечаются звездочками (от английского *star* — «звезда» и *gaze* — глазеть). Но мы пока посмотрим на то, как с помощью этой библиотеки выгрузить симпатичную табличку с описательными статистиками. Установим библиотеку и обратимся к ней (а заодно и к `tidyverse`):

```
install.packages("stargazer")
```

```
library(stargazer)
library(tidyverse)
```

Загрузим файл `food_coded.csv`, который содержит результаты опроса студентов колледжа, посвященного их пищевым привычкам (любимая еда, еда для восстановления душевного спокойствия, любимая национальная кухня и прочее).

```
food <- read.csv("https://allatambov.github.io/rprog/data/food_coded.csv")
```

Выберем с помощью функции `select()` из датафрейма `food` столбцы `calories_scone` и `calories_chicken` и выведем для них описательные статистики:

```
stargazer(dplyr::select(food,
                        c(calories_scone, calories_chicken)))
```

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: вс, май 03, 2020 - 23:11:11
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} & \multic
## \hline \\[-1.8ex]
## calories\_scone & 124 & 505.242 & 230.841 & 315.000 & 420.000 & 420.000 & 980.000 \\
## calories\_chicken & 125 & 577.320 & 131.214 & 265 & 430 & 720 & 720 \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

Вместо таблички мы получили много непонятного кода в консоли. Функция `stargazer()` по умолчанию возвращает код LaTeX для таблицы, этот код можно скопировать как есть, а можно сразу сохранить в файл, добавив аргумент `out = "mytex.tex"`, где "mytex.tex" — название файла. Обратите внимание: после сохранения в файле будет только код для таблицы, преамбулу и окружение для документа нужно будет добавлять самостоятельно.

Как быть тем, кто не использует LaTeX? Попросить `stargazer()` вывести код для таблицы в формате HTML и выгрузить его в файл с расширением `.htm`. По умолчанию такой файл будет открываться в браузере, но его можно открыть как обычный текстовый документ с помощью Word, Libre Office и

подобных программ.

```
stargazer(dplyr::select(food,
                        c(calories_scone, calories_chicken)),
          type = "html",
          out = "my_sum.htm")
```

Теперь можно найти файл `my_sum.htm` в рабочей папке и открыть с помощью Word или аналогичного редактора. Таблица выглядит так:

Таблица 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
calories_scone	124	505.242	230.841	315.000	420.000	420.000	980.000
calories_chicken	125	577.320	131.214	265	430	720	720

По умолчанию функция `stargazer()` возвращает число заполненных ячеек в столбце (`N`), среднее (`Mean`), стандартное отклонение (`St. Dev.`), минимум и максимум (`Min` и `Max`), нижний и верхний квартили (`Pctl(25)` и `Pctl(75)`).

Вносить изменения в этот файл, конечно, можно вручную, но все же кое-какие детали удобнее учесть еще до выгрузки — у функции `stargazer()` много аргументов. Например, округлим все числа до второго знака после запятой (аргумент `digits`), уберем из выдачи число наблюдений и квартили (аргумент `omit.summary.stat`) и добавим медиану (аргумент `median`).

```
stargazer(dplyr::select(food,
                        c(calories_scone, calories_chicken)),
          type = "html",
          out = "my_sum.htm",
          digits = 2,
          omit.summary.stat = c("n", "p25", "p75"),
          median = TRUE)
```

Получаем:

Таблица 2:

Statistic	Mean	St. Dev.	Min	Median	Max
calories_scone	505.24	230.84	315.00	420.00	980.00
calories_chicken	577.32	131.21	265	610	720

Можем добавить заголовок (аргумент `title`) и комментарии (аргумент `notes`), плюс, указать, что комментарии должны быть выровнены по правому краю (аргумент `notes.align`).

```
stargazer(dplyr::select(food,
                        c(calories_scone, calories_chicken)),
          type = "html",
          out = "my_sum.htm",
          digits = 2,
          omit.summary.stat = c("n", "p25", "p75"),
          median = TRUE,
          title = "Summary statistics",
          notes = "Source: Kaggle",
          notes.align = 'r')
```

Получаем:

Таблица 3: Summary statistics

Statistic	Mean	St. Dev.	Min	Median	Max
calories_scone	505.24	230.84	315.00	420.00	980.00
calories_chicken	577.32	131.21	265	610	720

Source: Kaggle

По умолчанию функция `stargazer()`, если не используется для выгрузки результатов регрессионной модели, принимает на вход датафрейм и «подготавливает» описательные статистики для него самостоятельно. Но при желании можно выгрузить и сам датафрейм как есть.

Для примера создадим датафрейм (точнее, это будет *tibble data.frame* от `tidyverse`) с описательными статистиками с группировкой по полу респондента:

```
my.tab <- food %>%
  group_by(Gender) %>%
  summarise(mean = round(mean(fruit_day), 2),
            sd = round(sd(fruit_day), 2))

my.tab
```

Теперь выгрузим его в файл, добавив аргумент `summary = FALSE` (выгружать как есть, не делать *summary*) и аргумент `rownames = FALSE`, чтобы скрыть номера строк в выдаче:

```
stargazer(my.tab, summary = FALSE,
          rownames = FALSE,
          type = "html",
          out = "food_summary.htm")
```

Получаем:

Таблица 4:

Gender	mean	sd
1	4.36	0.84
2	4.02	1.01

Если столбцы «склеиваются» из-за недостаточного расстояния между ними, можно поработать с аргументом `column.sep.width`.

Примечание. Чтобы добавить готовую таблицу с выдачей в html-файл или pdf-файл, созданный с помощью R Markdown (как в этом конспекте, без лишнего кода R и без кода HTML/LaTeX), нужно:

- убедиться, что в `type` в `stargazer()` указан `latex`, если генерим pdf-файл и `html`, если html-файл или файл Word;
- убедиться, что аргумента `out` для выгрузки в файл нет;
- в функции `stargazer()` добавить аргумент `header=FALSE`, чтобы убрать комментарии с ссылками на авторов пакета в начале;
- в разметке ячейки `{}` после `r` добавить опции `results='asis', echo=FALSE`.