Searching for the Sweet Spot for more sales in Parenting Magazines? Customers without a College degree are more likely to purchase Parenting Magazines.

Dina Abu Ghosh, Anusha Champaneria,Ahmed Lugya

22/03/2020

Abstract

We use Kid Creative dataset (StatsProf,2012) to analyze the relationship between having a college degree and sales of parenting magazines. We analyze other affecting factors such as Race and use Matching algorithms to observe how much impact having a college degree had on the purchase of parenting magazines. The purpose of this analysis is to help companies determine the type of magazine Ads that should be portrayed to customers. We identify that holding or not holding a college degree has a great impact on the purchase of parenting magazines.

Introduction

It is said that generally women read more magazines than men (Gayman, 2016) however, when it comes to parenting magazines specifically, the answer is almost unknown. With the world changing from men being the sole breadwinners and women staying at home, to both genders working full time and taking care of house duties equally, fatherhood magazines have gradually become very popular (Gayman, 2016). Nowadays, many people lack proper knowledge on parenting especially with a busy lifestyle and limited free time. It is convenient for full time workers to coordinate with their spouse when it comes to taking care of their child. Some also ask relatives with experience in taking care of children such as their parents, siblings, or grandparents to take care of the child while they are busy with work. Some also hire a babysitter if they can afford to do so, however, not everyone can afford it nor do they have relatives to help. Additionally there are also many single parents who try to manage taking care of their child along with house duties and a single income. In situations like this, many resort to informative books or magazines for parenting as a guidance.

Parenting magazines are generalized to all types of parents and are popularly used worldwide. They help parents guide their way into parenthood when it comes to their child's health, safety, behavior change etc. (Meredith, 2019). It further allows them to educate themselves on how to be good parents and important steps they need to take which they might not be aware of. Using matching techniques and linear regression, we observed that holding or not holding a college degree has a great impact on the purchase of parenting magazines.We also looked at a few assumptions a linear regression model must hold true in order for the results to be trusted.We concluded the report by offering a discussion on the ethics and limitations of the dataset we used and pointing out a few recommendations that could be of great importance to marketing agents.

Research question:

What impact does Race and holding a College degree have on investing in Parenting Magazines?

Data Set

The dataset Kid Creative dataset has a total of 14 attributes. Attribute HasCollege represents whether the person is getting a college education. It consists of two variables; 1 and 0 where 1 represents that the person has one or more years of college education and 0 represents no college education. Attribute White represents race where 1 means race is white and 0 means other. Lastly, attribute Prev_Parent _Mag represents if a parenting magazine was previously purchased where 1 means a parenting magazine was previously purchased and 0 means a parenting magazine was not previously purchased.

In our analysis we will compare attribute Prev_Parent_Mag with attribute HasCollege, and attribute white to see if Race,and education level determines if one has purchased a parenting magazine

Descriptive Analysis

Table 1 below shows an analysis of data from 673 customers. Customers who previously purchased parenting magazines were 38 compared to 485 customers who did not purchase parenting magazines. Customers of the white race were 316 compared to 207 customers who did not identify as white. Customers who had a college degree were 45 compared to 478 customers who had no college degree.

25 customers identifying as white purchased parenting magazines compared to 13 customers identifying as non-white and purchased parenting magazines. 4 customers holding a college degree purchased parenting magazines compared to 34 customers that hold no college degree and bought parenting magazines.

Table 1: Frequency of Race, College_educated and purchased Parenting Magazine

X	PurchasedParent_Magazine	Race	College_educated	Freq
1	No	Otherwise	No	153
2	Yes	Otherwise	No	9
3	No	White	No	291
4	Yes	White	No	25
5	No	Otherwise	Yes	41
6	Yes	Otherwise	Yes	4

Linear Regression

The next step is to use linear regression to determine the real impact of being College educated on purchases of parenting magazines. Our linear regression equation can be written in the form; PurchasedParenting_Magazines = b0 + b1* College_educated + Race.

```
===============================================
                        Dependent variable:
                    ---------------------------
                     PurchasedParent_Magazine
-----------------------------------------------
College_educated              0.044*
                              (0.024)

Race                           0.027
                              (0.023)

Constant                     0.053***
                              (0.020)

-----------------------------------------------
Observations                    673
R2                             0.008
Adjusted R2                    0.005
Residual Std. Error      0.278 (df = 670)
F Statistic            2.635* (df = 2; 670)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Figure 1:

The results in Figure 1 show that being College educated or not has a 4.4% effect on the purchase of parenting magazines where as Race has a 2.7% effect on the purchase of parenting magazines.

Model Validation

The results in figure 1 above provide us with the details needed to assess how well our model fits the data.

Residuals:

Residuals are distances between our obseruvtation and the model in use.A summary of the residuals can be used to come up with a conclusion of whether our model fit the data or not.We can not concluding using this measure as our data is entirely categorical.

F statistic:

Tha F statistic informs us about the relationship between the dependent and independent variables we are testing. Generally, a large F value indicates a stronger relationship.

Multiple R-squared:

Our R-squared is 0.008, and this is the measure of how close our data are to the linear regression model we are using. Since our R-squared is not closer to 1, we can conclude that this is not well fitting model.

Creating a pre-matching table (Unmatched)

In the next section we summarize the data using the covariates (Race and College educated). In particular, we shall check the balance in the dataset among the customers who purchased a parenting magazine(Treatment Group) and customers who did not purchase a parenting magazine(Control Group). We are interested in the mean and standard deviation of the covariates in the results. We are also interested in the Standardized Mean Differences (SMD) in the variables.

```
                            Stratified by PurchasedParent_Magazine
                             No           Yes          SMD
  n                           616           57
  College_educated = Yes (%)  172 (27.9)    23 (40.4)   0.264
  Race = White (%)            422 (68.5)    44 (77.2)   0.196
```

Figure 2:

Figure 2 above shows the summary statistics of the covariates (Race and College educated).The control group has 616 subjects and the Treatment group has 57 subjects. We can also observe the mean and standard deviations of these variables.

The last column shows the Standard Mean Differences (SMD). The SMD of holding a College degree is 0.264 and the SMD of Race being white is 0.196 both of which are greater than 0.1.This is a sign of imbalance in the dataset.Forexample variable Race has 422 subjects in the Control group versus 44 subjects in the Treatment group which is very big difference. Variable College educated has 172 subjects in the Control group versus 23 subjects in the Treatment group which is also a very big difference.Hence this is where we actually need to perform propensity score matching.

Matching

In the next section we shall proceed with matching. Using Greedy matching analysis, Sets of participants for the Treatment group and Control group are created by the matching algorithm where a matched set consists of at least one subject from the treatment group (i.e customers who purchased a parenting magazine) and one from the control group (i.e customers who did not purchase a parenting magazine with similar propensity scores.

```
                            Stratified by PurchasedParent_Magazine
                             No            Yes           SMD
  n                          11305         11305
  College_educated = Yes (%)  2653 (23.5)   2653 (23.5)  <0.001
  Race = White (%)            9764 (86.4)   9764 (86.4)  <0.001
```

Figure 3:

Figure 3 above shows the summary statistics after matching. First thing you might notice is that we mactched 11305 subjects in the control group to 11305 subjects in the Treatment group.If you look at the Standardised Mean Differences, they are very small,a sign that we have done a good job at matching.

Matching using nearest neighbor

In the next section We perform matching using the Nearest Neighbour method. In this method, a treated unit is matched to a control unit(s) that is closest in terms of distance measure such as a logit.

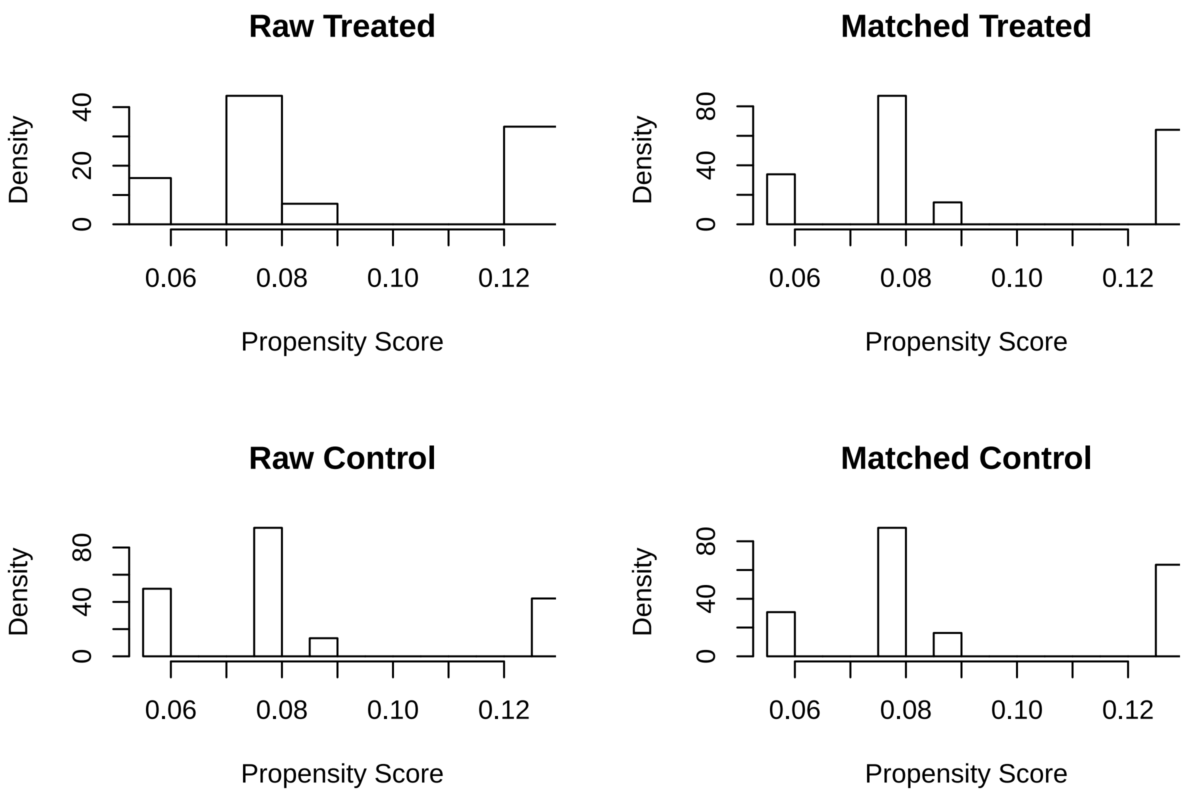

Figure 3:Histograms of Propensity Scores before and after matching.

Figure 3 above shows the histograms of our analysis before and after matching. The histograms on the left(before matching) differ to a great degree where as the histograms on the right(after matching) are significantly similar. Both the numerical and visual data show that the matching was done succefully.

Outcome analysis

In order to determine the causal difference, in the next section we shall carry out a paired t-test to test our hypothesis that there are no high purchases in parenting magazines when the customer has a college degree.

Before performing a paired t-test we create two subsets from the matched data above, one subset is for the Control group (y_con) and the other subset is for the Treatment group (y_trt). We use these two subsets to perform a pairwise difference.

One_Sample_t_test	**Values**
t	8.08
df	1.73e+04
p-value	6.95e-16
95% Confidence Interval(x)	0.0303
95% Confidence Interval(y)	0.0498
Sample estimates	0.04

Figure 4:

From Figure 4 above, we notice a p-value of 6.951e-16. This is a very small p-value and that makes the model highly significant.

Causal risk difference

The point estimate(mean) is 0.04, that is , the difference in probability of purchasing a parenting magazine when the customer has no college degree versus when the customer has a college degree is 0.04. This means that there are 4% higher chances of purchasing a parenting magazine when the customer has no college degree.

Ethics

Consent

The information collected in this study is very sensitive such as house hold income, marital status and home ownership. There is no mention of obtaining the information consensually, it fully states that each consumer has a "rich" profile and the information is gathered through third party sources. In which the data is purchased to be used in this study. This can indicate a major intrusive privacy invasion. The public is even more aware now of how their data should be protected especially with the recognition that scandals such as Facebook and Cambridge Analytica received publically. This has made users aware of their rights as consumers and how valuable their data is. It also shows how personal data can be monetized and exploited for profit and personal gain. Therefore, having a statement that says the information was obtained by purchasing it through a third-party source can be very problematic and can raise ethical concerns about the study as a whole and its results.

Attributes

The attributes seem to be either redundant or not fully comprehensive. For example, the race attributes states whether the user is "white" or "otherwise". It is inaccurate to group all non-white races into one attributes. That could reflect a large group of the sample and cannot be grouped into one category. In contrast to that, there are some attributes that are redundant and can cause confusion because there is no clear distinction between the two of them and can lead to inaccurate results. An example of that is the two following attributes used:

Employed in a Profession (IsProfessional = 1 if employed in a profession, 0 otherwise) Not employed (Unemployed = 1 if not employed, 0 otherwise)

The difference that the two attributes are supposed to highlight seems unclear. Having repetitive attributes can affect the results by giving s higher weight to the same attribute in comparison to other ones. Some very important information can be collected as it directly correlated to the research question such as number of kids a house hold has. None of the attributes covers number of kids, which can be vital to whether parents engage with these kind of advertisements, usually parents seek knowledge about parenting when they have their first child and by the time they have their second or third child they might be less interested in purchasing parenting magazine or subscribing to parenting websites as they are somewhat experienced at this point. So instead of having redundant attributes that lead to the same results, they can have a more comprehensive set of attributes that encompass more.

Statement on ethics

NeurIPS, which is a machine learning conference that asks for a statement on ethics with each submission(Rohan,2020) which essentially highlights the effects of the study and research conducted. When dealing with sensitive data such as the one in this dataset, a statement on ethics can help clarify the publics concerns and questions regarding the morality of the way such data was obtained.

APPENDIX

Source Code

This script has source code for all the analysis in Problem set 4.

```r
#### Contact details ####
# Title: Problem set 4
# Purpose: This script has source code for all the analysis in Problem set 4.
# Author: Dina, Anusha, Ahmed
# Last updated: 21 March 2020
# License: MIT License.

#### Setting up workspace ####
library(magrittr)
library(huxtable)
library(stargazer)
library(MatchIt)
library(ggplot2)
library(dplyr)
library(ddply)
library(janitor)
library(tidyverse)
library(plyr)
library(expss)
library(tableone)
library(knitr)
library(skimr)
library(papeR)
library(tidyr)
library(jtools)
library(sjmisc)
library(sjlabelled)
library(stargazer)
library(Matching)
library(GGally)


#Reading the raw data
purchase_data <- read.csv("http://logisticregressionanalysis.com/MiscPages/KidCreative.csv",
                          header=TRUE)

#Viewing the first few records in the dataset
head(purchase_data)

#After feedback

#Renaming columns to use
PurchasedParent_Magazine <- purchase_data$Prev.Parent.Mag
College_educated <- purchase_data$Has.College
Race <- purchase_data$White
Female <- purchase_data$Is.Female
```

```
#Creating a new dataset
mydata <- cbind(PurchasedParent_Magazine,Race,College_educated)
mydata <- data.frame(mydata)

#Apply labels
mydata = apply_labels(mydata,
                      PurchasedParent_Magazine = c("Yes" = 1,
                                                 "No" = 0),
                             Race = c("White" = 1,
                                           "Otherwise"=0),
                      College_educated = c("Yes" = 1, "No" = 0)
)

##Descriptive analysis
#Cross Tabulated table
xtb <- table(mydata)
#xtb
counts1 <- head(as.data.frame(xtb))
#Write to csv file count.csv
write.csv(x, "count1.csv")

#graph
ourplot <- ggpairs(purchase_data, columns = c("Prev.Parent.Mag", "Has.College", "White"),
                    columnLabels = c("PurchasedParent_Magazine", "College_educated", "white"))
ourplot

# Our covariates we using are:
xvars <- c("College_educated", "Race")

##Creating a Table 1, pre-matching
#look at table 1
table1 <- CreateTableOne(vars = xvars,strata = "PurchasedParent_Magazine", data = mydata,
                         test = FALSE)

#include Standardized Mean Differences
print(table1, smd = TRUE)

#Match
#do greedy matching on Mahalanobis distance
greedymatch <- Match(Tr=PurchasedParent_Magazine, M = 1, X=mydata[xvars])
matched <- mydata[unlist(greedymatch[c("index.treated","index.control")]),]

#createtableone for after matching
matchedtable1 <- CreateTableOne(vars = xvars,strata = "PurchasedParent_Magazine",data = matched,
                                test = FALSE)
print(matchedtable1, smd = TRUE)


#Outcome analysis
#If we want a causal difference, we can carry out a paired t-test

#Outcome analysis
y_trt <- matched$PurchasedParent_Magazine[matched$College_educated==1]
```

```
y_con <- matched$PurchasedParent_Magazine[matched$College_educated==0]

#pairwise difference
diffy <- y_trt - y_con

#paired t-test
#t.test(diffy)

#Print results table
httest <- hux(
  One_Sample_t-test = c("t", "df", "p-value", "95% CI","Sample estimates"),
  Values = c(8.0791,17303,6.951e-16,0.03033222/0.04976487,0.04004854),
  add_colnames = TRUE
)
right_padding(httest) <- 10
left_padding(httest)  <- 10
bold(httest)[1, ]          <- TRUE
bottom_border(httest)[1, ] <- 1
httest %>% set_background_color(everywhere, starts_with("S"), "orange")
# print_screen(httest)




#Frequency not in Cross-tab
# counts3 <- ddply(purchase_data, .(purchase_data$Has.College, purchase_data$Prev.Parent.Mag),
nrow)
# names(counts3) <- c("Has_College_degree","Previously_Purchased_Parent_Magazine", "Frequency")
# counts3

#Finding out the real impact of gender on the purchase
model1 <- lm(PurchasedParent_Magazine ~College_educated +
            Race,
          data= purchase_data)
model1

#summ(model1)
#Finding out the real impact of gender on the purchase
model1 <- lm(PurchasedParent_Magazine ~ College_educated + Race,
             data= purchase_data)
#model1

#Print results table
ht <- hux(
  Linear_regression_coefficients = c("(Intercept)", "College educated", "Race"),
  Values = c(0.05326,0.04390,0.02703),
  add_colnames = TRUE
)
right_padding(ht) <- 10
left_padding(ht)  <- 10
bold(ht)[1, ]          <- TRUE
bottom_border(ht)[1, ] <- 1
ht %>% set_background_color(everywhere, starts_with("S"), "orange")
```

```r
# print_screen(ht)
#stargazer(mydata)

#Use "comment=NA" to remove all hashes
#Use ```{r include=FALSE}
#knitr::opts_chunk$set(comment = NA)``` to remove all hashes globally

#Treatment group
Treatmentgroup <- subset(purchase_data,  Female == 1)
colMeans(Treatmentgroup)

#Control population
Controlgroup <- subset(purchase_data, Female == 0)
colMeans(Controlgroup)

#number of rows and columns we have in treatmentgroup
dim(Treatmentgroup)

#number of rows and columns we have in controlgroup
dim(Controlgroup)


#Propensity scores
pscores.model <- glm(PurchasedParent_Magazine ~ Female +
                Colleged_educated,
                family = binomial("logit"), data = purchase_data)
summary(pscores.model)

summ(pscores.model)

Propensity_scores <- pscores.model
purchase_data$PScores <- pscores.model$fitted.values
hist(purchase_data$PScores[Female==1],main = "PScores of Gender = 1")
hist(purchase_data$PScores[Female==0],main = "PScores of Gender = 0")

#Creating a Tableone pre-matching table
#covariates we are using
xvars <- c("Is.Female", "Has.College")
table1 <- CreateTableOne(vars = xvars,strata = "Prev.Parent.Mag",data = purchase_data, test
                         = FALSE)

print(table1, smd = TRUE)

#matching algorithms(Exact matching)
match1 <- matchit(pscores.model, method="exact",data=purchase_data)
summary(match1, covariates = T)

match1.data <- match.data(match1)
View(match1.data)

#create table one for exact matching
table_match1 <- CreateTableOne(vars = xvars,strata = "Is.Female",data = match1.data,test
                               = FALSE)
```

```r
print(table_match1, smd = TRUE)

#Nearest Neigbour Matching
m.out = matchit(Prev.Parent.Mag ~ Has.College + Race, data = purchase_data, method =
                  "nearest",
                ratio = 1)
#summary(m.out)
plot(m.out, type = "jitter")
plot(m.out, type = "hist")
```

References

- StatsProf (2012). What a Multivariate Logistic Regression Data Set Looks Like Retrieved from http://logisticregressionanalysis.com/303-what-a-logistic-regression-data-set-looks-like-an-example/

- Gayham, D. (2016). Nebraska Today. Study examines gender roles in parenting magazines.Retrieved from https://news.unl.edu/newsrooms/today/article/study-examines-gender-roles-in-parenting-magazines/

- Meredith, P. (2019). Parents. Parents Magazine. Retrieved from https://www.parents.com/parents-magazine/

- Thomas, D., Pastrana, S., Hutchings, A., Clayton, R. and Beresford, A. (2017). Ethical issues in research using datasets of illicit origin. Cambridge Cybercrime Centre, Computer Laboratory, p.2.

- Stefan Milton Bache and Hadley Wickham (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. https://CRAN.R-project.org/package=magrittr

- Jasjeet S. Sekhon. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R." Journal of Statistical Software, 42(7): 1-52.

- Daniel E. Ho, Kosuke Imai, Gary King, Elizabeth A. Stuart (2011). MatchIt:Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, Vol. 42, No. 8, pp. 1-28. URL http://www.jstatsoft.org/v42/i08/

- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York,

2016.

- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

- Sam Firke (2019). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 1.2.0. https://CRAN.R-project.org/package=janitor

- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

- Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

- Gregory Demin (2019). expss: Tables, Labels and Some Useful Functions from Spreadsheets and 'SPSS' Statistics. R package version 0.10.1. https://CRAN.R-project.org/package=expss

- Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer

- Kazuki Yoshida (2020). tableone: Create 'Table 1' to Describe Baseline Characteristics. R package version 0.11.1. https://CRAN.R-project.org/package=tableone

- Yihui Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.26.

- Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

- Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

- Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2019). skimr: Compact and Flexible Summaries of Data. R package version 2.0.2. https://CRAN.R-project.org/package=skimr

- B. Hofner (2019). papeR: A Toolbox for Writing Pretty Papers and Reports, R package version 1.0-4, https://CRAN.R-project.org/package=papeR.

- Hadley Wickham and Lionel Henry (2019). tidyr: Tidy Messy Data. R package version 1.0.0. https://CRAN.R-project.org/package=tidyr

- Long JA (2019)._jtools: Analysis and Presentation of Social Scientific Data_. R package version 2.0.1, <URL: https://cran.r-project.org/package=jtools>.

- Lüdecke D (2018). "sjmisc: Data and Variable Transformation Functions." *Journal of Open Source Software,3*(26), 754. doi: 10.21105/joss.00754 (URL: https://doi.org/10.21105/joss.00754).

- Lüdecke D (2020). *sjlabelled: Labelled Data Utility Functions (Version 1.1.3).* doi: 10.5281/zenodo.1249215 (URL: https://doi.org/10.5281/zenodo.1249215), <URL: https://CRAN.R-project.org/package=sjlabelled>.