# Logistic Regression: Script Lecture 4

Michaela Kreyenfeld (in collaboration with Cristina Samper)

Fall 2019

# 1 Review: Preparing your Data for an Investigation

In class we have been using data from the International Social Survey Programme ISSP. Appart from the yearly modules that we have used so far to investigate attitudes towards government spending and voting behavior, each year ISSP collects a set of demographic variables for analysis (https://www.gesis.org/issp/home/issp-background-variables (https://www.gesis.org/issp/home/issp-background-variables)). In this example we will investigate how religiousity is related to age and education. The aim is to review the data preparation process and show a nice way to display and compare the results of different models. In the end we will discuss the step-wise modeling stategy.

## 1.1 Getting an Overview of the Data

There are many strategies to explore the data with R, these ones are specifically for the case when the data is loaded with the `haven` package. One quick option to find a particular word within the variable labels is to use the function `lookfor(dataframe, "word")` from the `questionr` package. We will look for all variables available in the ISSP 2013 data related to attendance of religious events.

```
lookfor(ISSP2013, "attend")
```

```
##     variable                              label
## 171   ATTEND Attendance of religious services
```

Another solution to get an overview of all the variables in your data frame is to create a function that will create a list of the variables with their corresponding labels. In order to run this function you have to install and load the `tidyverse()` package. [1]

```
library(tidyverse)

makeVlist <- function(x) {
  labels <- sapply(x, function(x) attr(x, "label"))
  tibble(name = names(labels),
         label = labels)
}
```

Once you have created and run the function we can apply it to our data frame:

```
VariableLabels <- makeVlist(ISSP2013)

head(VariableLabels)
```

```
## # A tibble: 6 x 2
##   name     label
##   <chr>    <chr>
## 1 V1       GESIS Data Archive Study Number
## 2 V2       GESIS Archive Version
## 3 DOI      Digital Object Identifier
## 4 V3       Country/Sample (see V4 for codes for whole nation states)
## 5 V4       Country (see V3 for codes for the sample)
## 6 C_ALPHAN Country/ Sample Prefix ISO 3166 code - alphanumeric
```

# 1.2 Cleaning the Data

When we import the data using the `read_dta()` unless you specify otherwise the variables will be imported as ‘doubles’. In the case of categorical variables (factors), using the `attributes()` function we can see the labels that are attached to different integers. Then using `table()` we can see how many observations are in each category.

```
attributes(ISSP2013$ATTEND)
```

```
## $label
## [1] "Attendance of religious services"
## 
## $format.stata
## [1] "%8.0g"
## 
## $class
## [1] "haven_labelled"
## 
## $labels
##                  ES:NAP(Code 0 in ES_RELIG);TR:Not available
##                                                            0
## Several times a week or more often (incl. every day, several
##                                                            1
##                                                  Once a week
##                                                            2
##                                         2 or 3 times a month
##                                                            3
##                                                 Once a month
##                                                            4
##                                         Several times a year
##                                                            5
##                                                  Once a year
##                                                            6
##                             Less frequently than once a year
##                                                            7
##                                                        Never
##                                                            8
##                                                      Refused
##                                                           97
##                                                   Don't know
##                                                           98
##                                                    No answer
##                                                           99
```

```
table(ISSP2013$ATTEND)
```

```
## 
##     0     1     2     3     4     5     6     7     8    97    98    99 
##  1899  2055  4487  2855  2293  8058  4538  6289 11789   141   192   701
```

We only want to keep the observations that are valid for our dependent variable.

```
SUBSET01 <- subset(ISSP2013, ISSP2013$ATTEND<90 & ISSP2013$ATTEND >0 )
```

In this example we will also restrict our sample to valid answers in our independent variables. So we again use the functions above to see which values we want to keep

```
table(ISSP2013$AGE)
```

```
## 
##  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32 
##  32  98  97 550 710 656 693 667 722 711 762 703 746 750 764 814 768 820 
##  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50 
## 856 845 790 743 783 795 787 846 840 839 874 819 856 755 784 806 787 875 
##  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68 
## 784 828 811 832 830 785 755 798 763 761 714 727 767 742 727 627 607 580 
##  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86 
## 492 573 507 460 435 479 416 349 297 300 268 229 180 174 147 138 117 100 
##  87  88  89  90  91  92  93  94  95  96  97 112 999 
##  82  60  47  29  30  14  13   7   6   2   1   1 163
```

```
attributes(ISSP2013$DEGREE)
```

```
## $label
## [1] "Highest completed education level: Categories for international comparison"
## 
## $format.stata
## [1] "%8.0g"
## 
## $class
## [1] "haven_labelled"
## 
## $labels
##                                         No formal education 
##                                                            0 
##                                              Primary school 
##                                                            1 
## Lower secondary (secondary completed does not allow entry to 
##                                                            2 
##    Upper secondary (programs that allows entry to university 
##                                                            3 
## Post secondary, non-tertiary (other upper secondary programs 
##                                                            4 
## Lower level tertiary, first stage (also technical schools at 
##                                                            5 
##                       Upper level tertiary (Master, Doctor) 
##                                                            6 
##                                                   No answer 
##                                                            9
```

```
table(ISSP2013$DEGREE)
```

```
## 
##     0     1     2     3     4     5     6     9 
##  1879  3242 10350 11386  5779  7864  4414   383
```

Then we subset the valid answers:

```
SUBSET02 <- subset(SUBSET01, SUBSET01$AGE>=18 & SUBSET01$AGE<200 & SUBSET01$DEGREE<8)
```

Now let’s see how many respondents are in our final sample:

```
nrow(SUBSET02)
```

```
## [1] 41748
```

# 1.3 Variable Creation

Going back to the attibutes of the variable `ATTEND` I will define a person as religious if they attend religious services once a week:

```
SUBSET02$DEP <- -1
SUBSET02$DEP[SUBSET02$ATTEND==1 | SUBSET02$ATTEND==2] <- 1
SUBSET02$DEP[SUBSET02$ATTEND>2] <- 0
```

We will define a categorical variable for age and distinguish between ‘young’ and ‘old’.

```
SUBSET02$AGE_C <- "-1"
SUBSET02$AGE_C[SUBSET02$AGE>17] <- "1-young"
SUBSET02$AGE_C[SUBSET02$AGE>50] <- "2-old"
SUBSET02$AGE_C <- factor(SUBSET02$AGE_C)
```

For education we will distinguish between 3 categories.

```
SUBSET02$EDU_C <-"-1"
SUBSET02$EDU_C[SUBSET02$DEGREE==0 | SUBSET02$DEGREE==1 | SUBSET02$DEGREE==2] <-"1-low
"
SUBSET02$EDU_C[SUBSET02$DEGREE==3 | SUBSET02$DEGREE==4] <-"2-medium"
SUBSET02$EDU_C[SUBSET02$DEGREE==5 | SUBSET02$DEGREE==6] <-"3-high"
SUBSET02$EDU_C <- factor(SUBSET02$EDU_C)
```

# 1.4 Descriptive statistics of our sample

Before modeling the data it is always a good idea to look at your variables descriptively. In this manner you check whether they are well coded and functional, and you also grap an impression of the trends in the data.

By tabulating our dependent variable we can see what share of the people in our sample are religious:

```
# Table 1
addmargins(table(SUBSET02$DEP))
```

```
##
##     0     1   Sum
## 35339  6409 41748
```

From Table 1 I know `6409/41748=.1535163` around 15% of the respondents in the sample are religious

Now I can want to see how religiousity is related to age.

```
prop.table(table(SUBSET02$DEP, SUBSET02$AGE_C),2)
```

```
##
##       1-young     2-old
##   0 0.8564335 0.8338410
##   1 0.1435665 0.1661590
```

From this column percent table we can already see that older people are slighly more religious than younger people on average. Let's test whether this is a 'significant' difference by fitting a logistic regression.

# 1.5 Fitting a Logistic Regression Model

```
OUTPUT01 <- glm(DEP~ AGE_C, data=SUBSET02, family=binomial())
summary(OUTPUT01)
```

```
##
## Call:
## glm(formula = DEP ~ AGE_C, family = binomial(), data = SUBSET02)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6028  -0.6028  -0.5567  -0.5567   1.9703
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.78598    0.01866 -95.720  < 2e-16 ***
## AGE_C2-old   0.17288    0.02722   6.352 2.12e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35800  on 41747  degrees of freedom
## Residual deviance: 35760  on 41746  degrees of freedom
## AIC: 35764
##
## Number of Fisher Scoring iterations: 4
```

Now we can see that the difference is significant. Now to get a sense of the size of this difference we let's see what the odds ratios between old and young are, taking the young as our reference category.

```
exp(coef(OUTPUT01))
```

```
## (Intercept)  AGE_C2-old
##   0.1676329   1.1887249
```

Looking at our model results in the Odds Ratio scale we can see old people have 19% higher odds of being religious than younger people.

Now let’s see what happens when we control for education

```
OUTPUT02 <- glm(DEP~ AGE_C + EDU_C, data=SUBSET02, family=binomial())
summary(OUTPUT02)
```

```
## 
## Call:
## glm(formula = DEP ~ AGE_C + EDU_C, family = binomial(), data = SUBSET02)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6657  -0.6372  -0.5356  -0.5179   2.0373  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   -1.49116    0.02713 -54.969  < 2e-16 ***
## AGE_C2-old     0.09715    0.02781   3.494 0.000476 ***
## EDU_C2-medium -0.37806    0.03180 -11.890  < 2e-16 ***
## EDU_C3-high   -0.45006    0.03548 -12.683  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 35800  on 41747  degrees of freedom
## Residual deviance: 35553  on 41744  degrees of freedom
## AIC: 35561
## 
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(OUTPUT02))
```

```
##   (Intercept)    AGE_C2-old EDU_C2-medium   EDU_C3-high
##     0.2251108     1.1020256     0.6851869     0.6375898
```

We can now see adding the education variable changes the size of the effect of age on religiousity. In this model older respondents have only 10% higher odds than younger respondents of being religious.

Before we procede to interpert this change, let's review how to use the package `stargazer()` to display the models nicely side by side. Like this results are easy to compare and analyze. The package was however not designed specifically to report the results of a logistic regression so we have to make sure to specify the different statistics we want to display. In order to do this we have to extract information from the summaries of the models we have above. See below: first we extract the odds ratios, then we extract the confidence intervals and to end we also have to extract the significance measure. If we do not do this manually `stargazer()` will automatically calculate some of the wanted parameters (like the significance level) on its own.

```
# For OUTPUT01
OR.vector1 <- exp(coef(OUTPUT01))
CI.vector1 <- exp(confint(OUTPUT01))
```

```
## Waiting for profiling to be done...
```

```
p.values1 <- list(summary(OUTPUT01)$coefficients[,4])

# For OUTPUT02
OR.vector2 <- exp(coef(OUTPUT02))
CI.vector2 <- exp(confint(OUTPUT02))
```

```
## Waiting for profiling to be done...
```

```
p.values2 <- list(summary(OUTPUT02)$coefficients[,4])
```

Now we can insert all this information into the stargazer function: `coef` tells the function what coefficients to display; `ci=T` tells the function you want to see the confidence intervals displayed; `ci.custom=list(CI.vector1, CI.vector2)` feeds the confidence intervals we calculated in the form of a list; `single.row=F` means that you don't want to display the ci in the same line of the coefficients; `type="text"` is the format in which you want the table to be displayed (if you are creating a html document for example you would change this to `type="html"` ); `p=c(p.values1, p.values2)` is a list of the significance levels (note that we do not place list before the concatination of the p.values, this is because we already extracted them as a list above).

```
stargazer(OUTPUT01, OUTPUT02, coef = list(OR.vector1, OR.vector2), ci = T,
          ci.custom = list(CI.vector1, CI.vector2), single.row = F, type = "text", p=
c(p.values1, p.values2) )
```

See how the table looks in the section below.

# 2 Stepwise Modelling

Find below the results from the models above from the ISSP 2013. The way in which we have added one variable after the other is called step-wise modeling. It is a method that can help you understand the relationship between different covariates in the model.

To recap, the dependent variable is religiosity. As you can see from Model 1, older people are more religious. If one controls for education, the effect of religion is less strong. Thus, the correlation between age and religiosity is partially due to the fact that older people are less educated and that education is correlated with age. In this example, education was the **confounder**. If you control for a new variable and the effect of your variable of interest becomes stronger, one speaks of a **suppressor**.

```
stargazer(OUTPUT01, OUTPUT02, coef = list(OR.vector1, OR.vector2), ci = T,
          ci.custom = list(CI.vector1, CI.vector2), single.row = F, type = "html", p=
c(p.values1, p.values2) )
```

| | *Dependent variable:* | |
|---|---|---|
| | DEP | |
| | (1) | (2) |
| AGE_C2-old | 1.189$^{***}$ | 1.102$^{***}$ |
| | (1.127, 1.254) | (1.044, 1.164) |
| EDU_C2-medium | | 0.685$^{***}$ |
| | | (0.644, 0.729) |
| EDU_C3-high | | 0.638$^{***}$ |
| | | (0.595, 0.683) |
| Constant | 0.168$^{***}$ | 0.225$^{***}$ |
| | (0.162, 0.174) | (0.213, 0.237) |
| Observations | 41,748 | 41,748 |
| Log Likelihood | -17,879.760 | -17,776.420 |
| Akaike Inf. Crit. | 35,763.520 | 35,560.850 |
| *Note:* | | *p<0.1;* ***p<0.05;*** p<0.01 |

In this model we can also see that medium and high educated individuals are less religious than low educated individuals. In contrast to 'lowly' educated individuals 'medium' educated individuals have 32% decreased odds of being religious. Also 'highly' educated individuals have 36% decreased odds of being religious in comparison to 'lowly' educated individuals. Does this mean 'highly' educated individuals have lower odds than 'medium' educated individuals of being religious? From the model above we cannot really tell, we have only tested the differences with 'lowly' educated individuals. To find out whether the difference between the 'medium' educated and 'highly' educated individuals we have to calculate Model 2 again, but this time we will use 'medium' educated as our reference category.

```
SUBSET02$EDU_C <- relevel(SUBSET02$EDU_C, "2-medium")

OUTPUT03 <- glm(DEP~ AGE_C + EDU_C, data=SUBSET02, family=binomial())
```

To get the output for stargazer we have to again extract the information of the model as we did above:

```
OR.vector3 <- exp(coef(OUTPUT03))
CI.vector3 <- exp(confint(OUTPUT03))
```

```
## Waiting for profiling to be done...
```

```
p.values3 <- list(summary(OUTPUT03)$coefficients[,4])
```

Now we can look at the three models side by side:

```
stargazer(OUTPUT01, OUTPUT02, OUTPUT03, coef = list(OR.vector1, OR.vector2,OR.vector3
), ci = T, ci.custom = list(CI.vector1, CI.vector2,CI.vector3), single.row = F, type
= "html", p=c(p.values1, p.values2, p.values3), star.cutoffs = c(0.05, 0.01, 0.001) )
```

| | | *Dependent variable:* | |
|---|---|---|---|
| | | DEP | |
| | (1) | (2) | (3) |
| AGE_C2-old | 1.189$^{***}$ | 1.102$^{***}$ | 1.102$^{***}$ |
| | (1.127, 1.254) | (1.044, 1.164) | (1.044, 1.164) |
| EDU_C2-medium | | 0.685$^{***}$ | |
| | | (0.644, 0.729) | |
| EDU_C1-low | | | 1.459$^{***}$ |
| | | | (1.371, 1.553) |
| EDU_C3-high | | 0.638$^{***}$ | 0.931$^{*}$ |
| | | (0.595, 0.683) | (0.868, 0.998) |
| Constant | 0.168$^{***}$ | 0.225$^{***}$ | 0.154$^{***}$ |
| | (0.162, 0.174) | (0.213, 0.237) | (0.147, 0.162) |
| Observations | 41,748 | 41,748 | 41,748 |
| Log Likelihood | -17,879.760 | -17,776.420 | -17,776.420 |
| Akaike Inf. Crit. | 35,763.520 | 35,560.850 | 35,560.850 |
| *Note:* | | | *p<0.05;* **p<0.01;** p<0.001 |

Note that the difference between ‘medium’ and ‘highly’ educated individuals is weaker than the diffence between ‘medium’ and ‘lowly’ educated individuals. It is no longer significant at the 99% level, but only in the 95% level. I have also included the command `star.cutoffs = c(0.05, 0.01, 0.001)` in this table so the stars in the table match those displayed in `summary()`. In the tables displayed above (only with Models 1 and 2) the `*` corresponds to the 90% level, while `**` corresponds to 95% level. In this table `*` represents the 95% level.

[1] source: https://stackoverflow.com/tags/r-haven/hot?filter=all (https://stackoverflow.com/tags/r-haven/hot?filter=all)↩