Tutorial: Fitting GLMs to Survey Data

Marju Kaps (based on material from Andi Fugard)

3 Feb 2021

1 Introduction

By the end of this tutorial you will know how to:

- Fit GLMs on survey data.
- Perform model comparison on survey data.

This week, we are building on our previous tutorial on Survey Weights. Feel free to continue writing your code in the same file you used last week (in which case you can skip ahead to Section 3).

As we're not starting from scratch, let's load the same packages again.

```
library(dplyr)
library(ggplot2)
library(purrr)
library(survey)
library(srvyr)
```

2 Setup

We will continue to work with the 2011 Canadian National Election Study data (https://moodle.bbk.ac.uk/mod/folder/view.php?id=1108415).

Here are our 9 variables again:

Variable name	Description
`id`	household ID number.
`province`	a factor with (alphabetical) levels `AB`, `BC`, `MB`, `NB`, `NL`, `NS`, `ON`, `PE`, `QC`, `SK`; the sample was stratified by province.
`population`	population of the respondent's province, number of people over age 17.
`weight`	weight sample to size of population, taking into account unequal sampling probabilities by province and household size.
`gender`	a factor with levels `Female`, `Male`.
`abortion`	attitude toward abortion, a factor with levels `No`, `Yes`; answer to the question "Should abortion be banned?"
`importance`	importance of religion, a factor with (alphabetical) levels `not`, `notvery`, `somewhat`, `very`; answer to the question, "In your life, would you say that religion is very important, somewhat important, not very important, or not important at all?"
`education`	a factor with (alphabetical) levels `bachelors` (Bachelors degree), `college` (community college or technical school), `higher` (graduate degree), `HS` (high-school graduate), `lessHS` (less than high-school graduate), `somePS` (some post-secondary).
`urban`	place of residence, a factor with levels `rural`, `urban`.

Let's read in the data and check that all of our variables are as described.

```
ces <- read.csv("ces11.csv", stringsAsFactors = TRUE)
head(ces,10)
```

	id <int>	provin... <fct>	populati... <int>	weight <dbl>	gen... <fct>	aborti... <fct>	importa... <fct>	educati... <fct>
1	2851	BC	3267345	4287.85	Female	No	somewhat	somePS
2	521	QC	5996930	9230.78	Male	No	not	bachelors
3	2118	QC	5996930	6153.85	Male	Yes	somewhat	college
4	1815	NL	406455	3430.00	Female	No	very	somePS
5	1799	ON	9439960	8977.61	Male	No	not	higher

	id <int>	provin... <fct>	populati... <int>	weight <dbl>	gen... <fct>	aborti... <fct>	importa... <fct>	educati... <fct>
6	1103	ON	9439960	8977.61	Female	No	not	higher
7	957	NL	406455	3430.00	Female	Yes	very	lessHS
8	3431	NL	406455	1715.00	Female	Yes	notvery	college
9	2516	NL	406455	1715.00	Male	No	very	college
10	959	NL	406455	3430.00	Male	Yes	very	lessHS

1-10 of 10 rows

And we need to set up our survey object (using `srvyr`):

```
ces_s <- ces %>%
  as_survey(ids = id,
            strata = province,
            fpc = population,
            weights = weight)
```

And now we're all set up!

3 Fitting a GLM

The `survey` package makes this very easy. There is a command called `svyglm` which is identical to `glm` except it has a parameter called `design` instead of `data`.

See `?svyglm` for more information.

3.1 Activity | 3.2 Answer

a. `mutate` the survey object to add a binary variable called `againstAbortion` which is 1 if the participant is against abortion and 0 if not.
b. Fit an intercept-only logistic regression model without using weights (you can use `as_tibble` to get the "raw" data frame hidden within the survey object).
c. Do the same again, this time using the survey structure and weighting.
d. Transform the log odds outputted by your models into proportions. How do these predicted proportions compare with the proportions we calculated manually last time?

4 Trying out slopes

Now, having completed the traditional step of fitting an intercept-only model, we can give the slopes a go.

4.1 Activity | 4.2 Answer

Regress `againstAbortion` on `importance`, `education`, and `gender`, and interpret what you find.

5 Model comparison

We saw in the previous step that the respondent's gender, education level and importance of religion all seem to significantly predict views on abortion. Let's see if the model we just fitted gives us the best account of the data.

You wil need the `car` package for diagnostics.

```
library(car)
```

5.1 Activity | 5.2 Answer

a. Assess whether `gender` significantly increases the amount of variance accounted for by our GLM by removing it from the model computed in the previous section.
b. Try some diagnostics for the model that better accounts for the data.

6 Drawing generalisations

Depending on the goal of your analysis, you may want to play around with the way that your variables are set up in order to draw meaningful conclusions.

Does an increase in the level of education predict attitudes to abortion? Does an increase in the importance of religion predict attitudes to abortion?

Our previous analysis does actually not answer these questions directly since we compared different levels of the `education` and `importance` variables individually to a comparison level. The relevant variables were set up as factors, as we can see:

```
sapply(ces, class)
```

```
##         id   province population     weight     gender   abortion
importance
##  "integer"   "factor"  "integer"  "numeric"   "factor"   "factor"
"factor"
##  education      urban
##   "factor"   "factor"
```

If the variable forms a meaningful scale, we can convert it accordingly.

Let's look at the importance of religion as an example.

The levels for the `importance` variable are already in a meaningful order. Let's convert it to a scale using `as.numeric`.

We'll make it a scale from `0` (unimportant) to `3` (very).

(Note that this assumes that the possible answers to the survey question form a scale with equal increments - this may not be the case. How you set up or modify your variables depends on your assumptions.)

```
ces_s <- ces_s %>%
  mutate(importance.scale = as.numeric(ces$importance) - 1)
```

Check if it worked:

```
ftable(importance.scale ~ importance, data = ces_s)
```

```
##            importance.scale   0   1   2   3
## importance
## not                          607   0   0   0
## notvery                        0 315   0   0
## somewhat                       0   0 714   0
## very                           0   0   0 595
```

6.1 Activity | 6.2 Answer

Repeat the regression analysis with the three predictors (importance of religion, education level, and gender), this time using the converted version of the religion variable. What do you notice?

Discussion question: Which of the two versions of the importance of religion variable would *you* use in your analysis? Why? Try to think of pros and cons for each.