Fitting Regression Models to Data From Complex Surveys

An Appendix to *An R Companion to Applied Regression*, third edition

John Fox & Sanford Weisberg

last revision: 2018-09-09

Abstract

In this appendix, we illustrate how to use the **survey** package for R to represent the design of a complex survey sample and to fit a regression model to survey data.

Most of the statistical models fit to data in the *R Companion* assume independently sampled observations. Notable exceptions are the mixed-effects models discussed in Chapter 7, which are for clustered hierarchical or longitudinal data. In Section 7.2.7, we also discuss "sandwich" standard errors for least-squares regression coefficients that take clustering into account. In this appendix, we explain how to use the **survey** package for R to fit statistical models to data from complex sample surveys.

1 Basic Ideas

Real survey samples drawn from finite populations don't produce strictly independent observations. The simplest survey-sampling scheme used in practice is *simple random sampling* (*SRS*), in which n cases are drawn randomly from a complete list of N members of the target population, without replacement and with equal probability of selection at each step. Thus, all of the $\binom{N}{n}$ subsets of cases of size n have the same probability of selection. In contrast to *independent random sampling*, in which individuals are replaced in the population list before each subsequent selection, the observations in SRS are dependent, but, unless the *sampling fraction* n/N is large (say 5% or greater), the dependencies induced are trivial and can be ignored.

Most sample surveys use more complex sampling schemes than SRS. To fix ideas, we refer to the 2011 Canadian Election Study (CES) campaign-period survey.[1] Sampling for this survey proceeded as follows:

1. The country was divided into *strata* defined by the 10 Canadian provinces.[2]

2. Within each provincial stratum j, a simple random sample of n_j households was selected using a *sampling frame* of household phone numbers. Canadian provinces are very unequal in population size, and so to facilitate inter-provincial comparisons, smaller provinces were relatively over-sampled—that is, the strata sampling fractions n_j/N_j differed. As a consequence, different households had unequal, but known, probabilities of selection into the sample.

3. The sampled households were contacted by phone and the number of eligible voters in each household was determined. In a second stage of sampling, 1 respondent was selected randomly and with equal probability from among the eligible voters in the household. Thus, individuals

[1] The 2011 CES is described in detail in Fournier et al. (2013) and Northrup (2012).

[2] Because they are sparsely populated, voters residing in the three Canadian territories were not included in the target population of the survey.

in larger households had a smaller probability of selection into the sample than individuals in smaller households.

The eligible voters in a household constitute a *cluster*. If, at the final stage of sampling, had all the voters in each household been selected, then the survey would have employed *cluster sampling*. Cluster sampling typically induces non-trivial dependencies into the sample, making it less efficient (i.e., subject to greater sampling variation) than a simple random sample of the same size. In contrast, stratification generally produces samples that are modestly more efficient than SRS.

It's clear from the descriptions of the CES that each eligible voter in the population did not have an *equal* probability of selection into the sample. What's important, however, is that each individual have a *known* probability of selection. In analyzing the data, individuals in the sample are *weighted* in inverse proportion to their probability of selection, producing unbiased estimates of population characteristics. Weights may also be used to compensate for differential rates of global nonresponse to the survey (produced, e.g., by refusal to be interviewed or by failure to answer the phone). This procedure, which uses population data (e.g., from the Census) to adjust survey-sampling weights, is called *post-weighting*. The CES employed weights based on the differences in strata sampling fractions and household size.[3] Unequal probabilities of selection generally produce a sample that's less efficient than a similar size simple random sample for estimating population-wide characteristics, but may yield more precise estimates for certain comparisons (e.g., inter-provincial comparisons in the CES).

In *multistage sampling*, there is more than one step entailing random selection. As mentioned, the CES used a relatively simple two-stage sampling procedure, with households within strata sampled in the first stage, and individuals within households sampled in the second stage.

When the sampling design of a survey induces substantial dependencies among the sampled observations, as in cluster sampling, it's important to take the dependencies into account in estimating sampling variation, for example, in computing regression-coefficient standard errors. More generally, if the object of a study is to estimate characteristics of real, finite population—for example, the population of Canadian voters residing in the 10 provinces at a particular point during the 2011 election campaign—then the sampling design should be taken into account. Had we access to the whole population, and therefore had the ability to fit a regression model to the population, then the parameters of the model would be known. Inference about these real-population parameters is termed *design-based inference*. In contrast, if the object of inference is the *process* that gave rise to the a real population, then even the coefficients of a model fit to the whole population are subject to statistical uncertainty, leading to what's termed *model-based inference*.

2 An Example: Attitudes Toward Abortion in the 2011 CES

Data drawn from the 2011 CES survey are available in the data set `CES11` from the **carData** package:[4]

```
library("survey")
 Loading required package: grid
 Loading required package: Matrix
 Loading required package: survival

 Attaching package: 'survey'
```

[3] Sampling weights are to be distinguished from inverse-variance weights, used in weighted-least-squares regression to adjust for nonconstant error variance; for WLS regression, see Section 4.9.4 of the *R Companion*.

[4] This example is borrowed from Fox (2016, Sec. 15.5).

```
 The following object is masked from 'package:graphics':

     dotchart

library("car")

 Loading required package: carData

brief(CES11)

 2231 x 9 data.frame (2226 rows omitted)
        id province population  weight gender abortion importance education urban
       [i]      [f]        [i]     [n]    [f]      [f]        [f]       [f]   [f]
 1    2851      BC     3267345 4287.85 Female      No    somewhat somePS    urban
 2     521      QC     5996930 9230.78 Male        No    not      bachelors urban
 3    2118      QC     5996930 6153.85 Male        Yes   somewhat college   urban
 . . .
 2230 2488      BC     3267345 4287.85 Female      No    not      higher    urban
 2231 1368      MB      871460 5829.16 Male        No    not      HS        urban

CES11$education <- factor(CES11$education, levels=c("lessHS", "HS", "somePS",
                            "college", "bachelors", "higher"))
```

In addition to reading the data, we load the **survey** and **car** packages. We'll use several functions in the latter, and we need the former to define the sampling design and to fit a statistical model to the data.

The variables in the `CES11` data set are as follows:

`id` household ID number; were more than one individual sampled per household, this variable would define clusters.

`province` standard two-character abbreviations for the 10 Canadian provinces, which define the strata of the sampling design.

`population` the population size of the province in which each respondent resides, used to compute the *finite population correction* for coefficient estimates; because the sampling fractions in the strata are very small, these corrections are negligble.

`weight` the sampling weight for each case, inversely proportional to the case's probability of selection; the weights are scaled so that they can be used to estimate population counts and totals, but the scaling doesn't affect the estimates that we report below.

`gender` a factor with levels `"Female"` and `"Male"`.

`abortion` a factor derived from the question, "Should abortion be banned?" with levels `"No"` and `"Yes"`.

`importance` a factor derived from the question, "In your life, would you say that religion is very important, somewhat important, not very important, or not important at all?" asked only of respondents who reported that they had a religion; respondents with no religion were assigned the response "not important at all"; the levels of the factor are coded `"not"`, `"notvery"`, `"somewhat"`, and `"very"`.

`education` a factor with levels `"lessHS"` (less than high school), `"HS"` (high-school graduate), `"somePS"` (some post-secondary), `"college"` (degree from a community college or technical institute), `"bachelors"` (Bachelor's degree), and `"higher"` (Master's or Doctorate); we redefine the factor to put its levels in their natural (i.e., non-alphabetical) order.

`urban` a factor with levels `"rural"` and `"urban"`.

Prior to analyzing the data we must establish the survey sampling design, via the `svydesign()` function in the **survey** package, producing an object that includes the `CES11` data frame along with information about the survey design:

```
CES.svy <- svydesign(ids=~id, strata=~province, fpc=~population,
                     weights=~weight, data=CES11)
```

The `svydesign()` function, and the **survey** package more generally, are very powerful and general, accomodating many different kinds of designs for complex surveys. The **survey** package is extensively documented by its author in Lumley (2010), which is slightly out-of-date but is supplemented by vignettes in the package.[5] We used several arguments in the call to `svydesign()`, not all of which are strictly required:

`ids` a one-sided formula that defines the clusters in the sampling design. Because theere's only 1 observation per cluster in the CES data set, we could have specified `ids=~0` (see below), indicating no clustering. In a more complex sampling design there may be clusters selected a several stages.

`strata` a formula defining the strata, in our case by `province`.

`fpc` a formula with information for computing the finite population correction, given here as the `population` size of each respondent's stratum; because the sample in each province is a tiny fraction of the population, this argument could be omitted with virtually no effect.

`weights` a formula specifying the sampling weights for the cases, given by the variable `weight` in the `CES11` data set.

`data` the data frame containing the data, `CES11`.

We wish to perform a binary logistic regression of attitude toward `abortion` on the survey respondents' `gender`, level of `education`, `urban` versus rural residence, and the `importance` attached to religion. First, however, let's examine the distribution of the response factor, `abortion`, in two ways: from the original data set, which effectively ignores the sampling design, and using the survey-design object that we created:

```
prop.table(xtabs(~abortion, data=CES11))
```

```
abortion
       No       Yes
0.8148812 0.1851188
```

```
svymean(~abortion, design=CES.svy)
```

```
               mean     SE
abortionNo  0.81502 0.0099
abortionYes 0.18498 0.0099
```

The results in this case are very similar: About 81.5% of respondents oppose banning abortion. Perhaps counterintuitively, we use the `svymean()` function from the **survey** package to compute the design-based estimate of the distribution of the factor `abortion`, which generates 0/1 dummy variables for each of the two categories of the factor. The mean of a 0/1 variable is the proprtion of 1s. The `svymean()` function also reports the standard error of each proportion. The standard error of a proportion p in an independent random sample is $\mathrm{SE}(p) = \sqrt{p(1-p)/n}$.

[5] See `vignette(package="survey")` for a list of vignettes, which can then be referenced by name—e.g., `vignette("survey")`; also see `help("svydesign")`.

We fit the logistic-regression model with the `svyglm()` function from the **survey** package.[6] A slight wrinkle is that we must use the `quasibinomial` rather than the `binomial` family to avoid a warning about noninteger counts produced by the use of differential sampling weights (reader: try it!):

```
mod.abortion.svy <- svyglm(abortion ~ importance + gender  + education + urban,
                           design=CES.svy, family=quasibinomial)
summary(mod.abortion.svy)

 Call:
 svyglm(formula = abortion ~ importance + gender + education +
     urban, design = CES.svy, family = quasibinomial)

 Survey design:
 svydesign(ids = ~id, strata = ~province, fpc = ~population, weights = ~weight,
     data = CES11)

 Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
 (Intercept)         -2.5977     0.3223  -8.059 1.25e-15
 importancenotvery    0.4579     0.3478   1.317  0.18812
 importancesomewhat   1.3265     0.2714   4.887 1.10e-06
 importancevery       3.1414     0.2618  12.000  < 2e-16
 genderMale           0.3280     0.1483   2.211  0.02712
 educationHS         -0.4446     0.2385  -1.864  0.06240
 educationsomePS     -0.8521     0.2900  -2.938  0.00334
 educationcollege    -0.5623     0.2395  -2.348  0.01897
 educationbachelors  -0.9802     0.2501  -3.919 9.15e-05
 educationhigher     -0.6754     0.3089  -2.187  0.02887
 urbanurban          -0.2830     0.1661  -1.704  0.08861

 (Dispersion parameter for quasibinomial family taken to be 0.9724821)

 Number of Fisher Scoring iterations: 5
Anova(mod.abortion.svy)
 Analysis of Deviance Table (Type II tests)

 Response: abortion
            Df    Chisq Pr(>Chisq)
 importance  3 253.6117  < 2.2e-16
 gender      1   4.8899   0.027014
 education   5  17.8308   0.003166
 urban       1   2.9019   0.088474
```

The `Anova()` function in the **car** package, along with `deltaMethod()` and `linearHypothesis()`, knows how to handle `"svyglm"` objects, producing by default type-II Wald tests.[7]

Let's compare these design-based results to the the model-based results obtained by using `glm()` to fit the logistic regression to the original `CES11` data set:

```
mod.abortion <- glm(abortion ~ importance + gender  + education + urban, data=CES11,
                    family=binomial)
compareCoefs(mod.abortion.svy, mod.abortion)
```

[6] See `apropos("^svy*")` and the correspond help files for other functions that may be applied to survey-design objects.

[7] Although we don't require it here, the **effects** package also includes methods for displaying models fit by `svyglm()`.

```
Calls:
1: svyglm(formula = abortion ~ importance + gender + education + urban, design = 
  CES.svy, family = quasibinomial)
2: glm(formula = abortion ~ importance + gender + education + urban, family = 
  binomial, data = CES11)

                   Model 1 Model 2
(Intercept)         -2.598  -2.545
SE                   0.322   0.280

importancenotvery    0.458   0.442
SE                   0.348   0.310

importancesomewhat   1.327   1.203
SE                   0.271   0.235

importancevery       3.141   2.977
SE                   0.262   0.225

genderMale           0.328   0.375
SE                   0.148   0.127

educationHS         -0.445  -0.322
SE                   0.238   0.194

educationsomePS     -0.852  -0.651
SE                   0.290   0.235

educationcollege    -0.562  -0.508
SE                   0.240   0.199

educationbachelors  -0.980  -0.901
SE                   0.250   0.208

educationhigher     -0.675  -0.937
SE                   0.309   0.266

urbanurban          -0.283  -0.306
SE                   0.166   0.136
```

```
Anova(mod.abortion, test.statistic="Wald")
 Analysis of Deviance Table (Type II tests)

 Response: abortion
            Df    Chisq Pr(>Chisq)
 importance  3 311.3160  < 2.2e-16
 gender      1   8.6737   0.003228
 education   5  25.2797   0.000123
 urban       1   5.0844   0.024142
```

Here, for greater comparability, we specify the argument `test.statistic="Wald"` to `Anova()` to get type-II Wald statistics rather than the default likelihood-ratio tests for a GLM. The pattern of results for the design-based and model-based estimates are similar: Holding the other predictors constant, opposition to abortion increases with the importance of religion; is greater for males than for females, is generally higher at lower levels of education (although the relationship to education isn't monotone), and is lower among urban than rural residents. The coefficient standard errors for the design-based estimates are larger, however, reflecting the unequal probabilities of selection for the respondents, and, partly as a consequence, the *p*-values for the tests of the various terms in the model are larger for the design-based estimates.

Finally, we verify that with one respondent per household cluster and very small sampling fractions for the provincial strata, it isn't really necessary to specify the `ids` and `fpc` arguments in defining the survey design:

```
CES.svy.2 <- svydesign(ids=~0, strata=~province, weights=~weight, data=CES11)
mod.abortion.svy.2 <- svyglm(abortion ~ importance + gender  + education + urban,
                              design=CES.svy.2, family=quasibinomial)
compareCoefs(mod.abortion.svy, mod.abortion.svy.2)
```

```
Calls:
1: svyglm(formula = abortion ~ importance + gender + education + urban, design =
  CES.svy, family = quasibinomial)
2: svyglm(formula = abortion ~ importance + gender + education + urban, design =
  CES.svy.2, family = quasibinomial)

                   Model 1 Model 2
(Intercept)         -2.598  -2.598
SE                   0.322   0.322

importancenotvery    0.458   0.458
SE                   0.348   0.348

importancesomewhat   1.327   1.327
SE                   0.271   0.271

importancevery       3.141   3.141
SE                   0.262   0.262

genderMale           0.328   0.328
SE                   0.148   0.148

educationHS         -0.445  -0.445
SE                   0.238   0.238

educationsomePS     -0.852  -0.852
SE                   0.290   0.290

educationcollege    -0.562  -0.562
SE                   0.240   0.240

educationbachelors   -0.98   -0.98
SE                    0.25    0.25

educationhigher     -0.675  -0.675
SE                   0.309   0.309

urbanurban          -0.283  -0.283
SE                   0.166   0.166
```

3 Complementary Reading and References

There is a vast literature on conducting and analyzing data from sample surveys. We mention only a few key references here.

- Lumley (2010) is the definitive reference for the **survey** package, and it also provides a good, accessible general introduction to survey sampling and estimation.
- Groves et al. (2011) is a wide-ranging introductory text on survey methods, including sampling.
- Fuller (2009) provides an extensive, if technical, treatment of survey sampling and estimation.

References

Fournier, P., Cutler, F., Soroka, S., and Stolle, D. (2013). Canadian Election Study 2011: Study documentation. Technical report, Canadian Opinion Research Archive, Queen's University, Kingson, Ontario.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models.* Sage, Thousand Oaks CA, third edition.

Fuller, W. A. (2009). *Sampling Statistics.* Wiley, Hoboken NJ.

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey Methodology.* Wiley, Hoboken NJ, second edition.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* John Wiley & Sons, Hoboken, NJ.

Northrup, D. (2012). The 2011 Canadian Election Survey: Technical documention. Technical report, Institute for Social Research, York University, Toronto, Ontario.