EXERCISE ON WEIGHTED DATA

This exercise uses the data from the Gelman et al BDA book web site. The data are available in the directory with this exercise on the 515 Course web site, in both the original `.dta` file format and in `.RData` format.

There are two data sets. One is data from a poll of voting preferences of individuals in a sample, with preferences for Bush indicated by a 1 and for others by a zero. For this variable, but (I think) not others, there are NA values that have to be handled — for this exercise, by dropping those observations. The variables on the polls data set include, besides `bush`, individual characteristics, state of residence (as a number) and `weight`. The `weight` variable is meant to reflect over or under sampling, so that weighted averages using these weights estimate population means.

The census data set has the same set of individual characteristics as the polls set, plus a variable `N` that is the number of people with the given characteristics (including state of residence).

We consider two ways to estimate the population proportion of people who prefer Bush. The obvious one is to take a weighted average, using the `weight` variable, The other is to follow the BDA discussion of these data and fit a logit model explaining the `bush` variable as a function of individual characteristics, then use the census numbers to weight the logit predictions for each type in the `census88` data set, making no use of the `weight` data at all.

(1) Estimate the overall population proportion favoring Bush, using both methods. For the logit model, use all the available characteristics, treating age, education, and state as factors. That is, rather than enter these variables into the prediction matrix as integers, you should break them into sets of dummy variables. In R this can be done using the **`as.factor`** function.

Do the same thing for state number 29 by itself.

(2) For each of the four estimates you have constructed, calculate a measure of uncertainty. For the weighted averages, use frequentist asymptotic distribution theory to get a standard error. assume the data are iid, and iid conditional on state 29 in that calculation. For the logit estimators, begin by using MCMC to sample from the posterior on the logit coefficients, then use those results to arrive at posterior standard errors for the estimates.

For a pure Horwitz-Thompson estimator the mean of the weighted Y's is an unbiased estimate of the object of interest, without any normalization of the weights to sum to one. Here, though, since we have no count of "unselected" observations, we do need to divide by the sum of the weights, and this contributes to the sampling error in the estimate.

Date: November 12, 2018.

©2018 by Christopher A. Sims. ©2018. This document is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
http://creativecommons.org/licenses/by-nc-sa/3.0/

(3) Explain why the validity of the logit approach, ignoring weights, depends on the weights being uncorrelated with the `bush` variable conditional on the observed individual characteristics. Is there a way to check whether this seems to be true in this sample?

(4) There are some states with no observations in the polls data, as well as one with data, but with all the `bush` values the same — i.e. sample standard deviation zero. Does this call into question the frequentist distribution theory for the weighted-data estimates? It seems to be possible to present estimates for these states based on the logit model. How is this possible? Should we be suspicious of the logit estimates for these states?

(5) Suppose the only data you had was for `bush`, `weight`, and `state`, with no census data at all. How might you proceed with a Bayesian approach?

(6) My impression is that the only characteristic needed to get rid of selection bias, i.e. the only one you have to condition on to get indepencence between `bush` and `weight`, is `black`. Can you see a way to check this? If this were true, would it be better to omit the other variables from the logit model?

Suggestions: You can easily estimate the logit model using the R **`glm`** function with a logit link function family, but you need a program to evaluate the binomial logit model's likelihood for the next part of the exercise, and you could also use that to search for the flat-prior posterior mode. You also need an estimate of the inverse Hessian matrix to use for constructing an MCMC jump distribution, which an optimization program will provide. R programs to evaluate the logit likelihood and to evaluate the agrregate and state level logit predictions are available on the course web site with the other material for this exercise.