Weighting

**Weights in statistics**

Posted by [Andrew](https://statmodeling.stat.columbia.edu/author/andrew/) on 17 January 2021, 9:23 am

Thomas Lumley [writes](https://notstatschat.rbind.io/2020/08/04/weights-in-statistics/):

There are roughly three and half distinct uses of the term weights in statistical methodology, and it’s a problem for software documentation and software development. Here, I [Lumley] want to distinguish the different uses and clarify when the differences are a problem. I also want to talk about the settings where we know how to use these sorts of weights, and the ones where we don’t. . . .

You can read the whole thing at the above link.

I agree with Lumley!

Weighting causes no end of confusion both in applied and theoretical statistics. People just assume because something has one name (“weights”), it is one thing. So then we get questions like, “How do you do weighted regression in Stan,” and we have to reply, “What is it that you actually want to do?”

And then there’s this whole thing where people do poststratification weighting and think they’re doing inverse probability weighting; see Section 3.3 of [this article with John Carlin](http://www.stat.columbia.edu/~gelman/research/published/handbook5.pdf) to see why these two sorts of weights are different.

Here’s what we wrote about weighting in Section 10.8 of Regression and Other Stories:

**Three models leading to weighted regression**

Weighted least squares can be derived from three different models:

1. *Using observed data to represent a larger population*. This is the most common way that regression weights are used in practice. A weighted regression is fit to sample data in order to estimate the (unweighted) linear model that would be obtained if it could be fit to the entire population. For example, suppose our data come from a survey that oversamples older white women, and we are interested in estimating the population regression. Then we would assign to survey respondent a weight that is proportional to the number of people of that type in the population represented by that person in the sample. In this example, men, younger people, and members of ethnic minorities would have higher weights. Including these weights in the regression is a way to approximately minimize the sum of squared errors with respect to the population rather than the sample.

2. *Duplicate observations*. More directly, suppose each data point can represent one or more actual observations, so that i represents a collection of w\_i data points, all of which happen to have x\_i as their vector of predictors, and where y\_i is the average of the corresponding wi outcome variables. Then weighted regression on the compressed dataset, (x, y, w), is equivalent to unweighted regression on the original data.

3. *Unequal variances*. From a completely different direction, weighted least squares is the maximum likelihood estimate for the regression model with independent normally distributed errors with unequal variances, where sd(ε\_i) is proportional to 1/√w\_i . That is, measurements with higher variance get lower weight when fitting the model. As discussed further in Section 11.1, unequal variances are not typically a major issue for the goal of estimating regression coefficients, but they become more important when making predictions about individual cases.

These three models all result in the same point estimate but imply different standard errors and different predictive distributions. For the most usual scenario in which the weights are used to adjust for differences between sample and population, once the weights have been constructed and made available to us, we first renormalize the vector of weights to have mean 1 (in R, we set w <- w/mean(w)), and then we can include them as an argument in the regression (for example, stan\_glm(y ~ x, data=data, weights=w)).

**4**