Statistical Science
2017, Vol. 32, No. 2, 265–278
DOI: 10.1214/16-STS605

Fitting Regression Models to Survey Data

Thomas Lumley and Alastair Scott

Abstract. Data from complex surveys are being used increasingly to build the same sort of explanatory and predictive models used in the rest of statistics. Although the assumptions underlying standard statistical methods are not even approximately valid for most survey data, analogues of most of the features of standard regression packages are now available for use with survey data. We review recent developments in the field and illustrate their use on data from NHANES.

Key words and phrases: Complex sampling, statistical graphics.

1. INTRODUCTION

The topic of regression modelling of data from complex samples contains several quite different research worlds. The largest of these worlds is secondary analysis of public-use data from large in-person or telephone surveys, conducted by people who are experts in the subject matter rather than in sampling. This is also where implementations are most widely available and whether methodology is tidiest and most complete, though sometimes at the cost of unrealistic assumptions about nonresponse. In this paper, sampling-weighted regression analysis of large public-use datasets will be the core theme, but we will also describe limitations of this approach and situations where it is possible to do better.

As a concrete example, we will often refer to public-use data from the National Health And Nutrition Examination Surveys (NHANES) conducted by the US National Center for Health Statistics (National Center for Health Statistics, 1994). These are a series of large, multistage samples of the US civilian, noninstitutionalised population, which involve clinical examinations as well as detailed interviews. For practical reasons, the surveys have highly stratified multistage sampling with only a small number of city- or county-level sampling units at the first stage. The public use data do not contain all the variables used to design the sample, and in fact present a simplified version of the design as a two-stage sample with stratification and clustering only at the first stage. In this simplified version of the design, there are 30 clusters in each two-year sampling wave, chosen in pairs from 15 strata. The number of participants per cluster ranges from 500 to 900. People under 18 and over 60 are oversampled, as are Mexican Americans, African-Americans and low-income White Americans. Participants complete a health and diet interview and clinical examination including blood draw. NHANES has been a valuable resource in medical and public health research, and tens of thousands of papers have been published using the data. The data and R code we used are available at https://github.com/tslumley/regression-paper.

We shall suppose throughout that we are given a set of observations $\{(y_i, \boldsymbol{x}_i); i \in \mathcal{S}\}$ on a response variable, y, and a vector of possible explanatory variables, $\boldsymbol{x}$, together with associated weights, $\{w_i; i \in \mathcal{S}\}$, from a sample, $\mathcal{S}$, of n units drawn from a finite population or cohort of N units. Broadly speaking, w_i is an indication of the number of population units represented by the ith sample unit. In some cases, w_i will be equal to $1/\pi_i$, where π_i is the probability of selecting the ith unit under some probability sampling design; more often, w_i will be adjusted using post-stratification or raking to match known population totals to compensate for nonresponse and frame errors. We write R_i for the indicator that unit i in the population was sampled.

We also assume we are given enough information about the design to estimate variances—typically either stratum and primary sampling unit (PSU, "cluster") identifiers or sets of resampling ("replicate") weights.

For most of the paper, we will consider marginal generalised linear models. We assume that the popu-

Thomas Lumley is Professor of Biostatistics and Alastair Scott is Emeritus Professor, Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand (e-mail: t.lumley@auckland.ac.nz; a.scott@auckland.ac.nz).

lation is a realisation of some probability model with density $f(Y|X;\beta)$ (the "superpopulation" model) in which

$$g\big(E[Y|X=x]\big)=g(\mu)=\eta=\boldsymbol{x}'\beta \tag{1.1}$$

with $\dim(\beta)=p$, and in which the marginal variance can reasonably be approximated by

$$\operatorname{var}[Y|X=x]=\sigma^2V(\mu). \tag{1.2}$$

We will refer to the exponential family model with these means and variances as the "working model". In some places, we will additionally assume that the working model gives the true marginal distributions of $Y|X$; in other places, we will consider the possibility that even the mean model might be misspecified. We will write $E_\pi[\cdot]$ for expectations over the finite-population sampling, $E_P[\cdot]$ for expectations over the model, and $E[\cdot]$ for expectations over both.

In some settings, we will need to consider asymptotics. When the primary interest is in the marginal regressions there does not seem to be any important loss of generality in treating the population as an i.i.d. sample of (X,Y). The correlation structure of the data can be treated as the result of auxiliary variables such as latitude and longitude that are part of the sampled vector but not of interest in the model, or it can be regarded as purely an artefact of sampling, with the population being sorted into strata and clusters after it is created (Lumley and Scott, 2013). We always assume that $n\to\infty$ and $n/N\to c\in[0,1)$, and will need to make additional assumptions about clusters and strata in specific contexts. We will write β_0 for the true parameter value in the superpopulation model, $\tilde{\beta}_N$ for the maximum quasi-likelihood estimator of β_0 that would be obtained from full population data (the "census parameter"), β^* for the limit in probability of $\tilde{\beta}_N$ and $\hat{\beta}_n$ for the maximum pseudo-likelihood estimator to be described in Section 3.

We note for future research that it would be valuable to have sampling asymptotics better founded in the spatial structure of populations, not only for a better match to reality but also because it could simplify the development of Donsker-type theorems, uniform tail bounds, and other machinery of modern mathematical statistics.

The layout of the paper is as follows: in Section 2, we describe how to extend familiar exploratory analyses to survey data; in Section 3 we consider pseudo-likelihood estimators based on the working likelihood and whether the weights are necessary or desirable; in Section 4 we describe tests and model selection criteria based directly on the working likelihood; in Section 5 we consider other ways to use the weights more efficiently; and in Section 6 we give a brief overview of situations where true maximum likelihood estimation is possible.

2. EXPLORATORY ANALYSIS

Exploratory data analysis is every bit as important for regression on survey data as in any other context. Simple data summaries are easy to extend to use sampling weights; here we restrict our attention to scatterplots and smoothers, where the appropriate extensions are less obvious.

Korn and Graubard (1998) described approaches to both problems. For scatterplots, they give two recommendations. One, which is useful for relatively small data sets, is to scale the plotting symbol to have area proportional to the weight for an observation. This gives so-called "bubble plots". Their other recommended approach, most useful for large data sets, is "thinning" the data by sampling $m\ll n$ observations with the probability for observation i proportional to w_i (and so inversely proportional to π_i). The resulting subsample is an equal-probability, though not independent, sample from the population and an ordinary scatterplot can be drawn. As the resulting scatterplot is random, it is usually recommended to take more than one replicate of the subsample and confirm that visually important patterns persist.

Two further methods are described by Lumley (2010), using familiar techniques for scatterplots of large data sets (Unwin, Theus and Hofmann, 2007). In alpha-blending, points are partially transparent, with opacity proportional to w_i. When points are overplotted, the opacity accumulates and the result is, essentially, a two-dimensional density estimate for the population bivariate distribution. In hexagonal binning (Carr et al., 1987), the plotting area is partitioned into a grid of hexagons. The total weight for observations in each cell is computed, and hexagons drawn in each cell with area proportional to the total weight.

Figure 1 shows the relationship between diastolic blood pressure and age in the 13,957 individuals from the 2003–2004 and 2004–2005 waves of NHANES who participated in the clinical exam and dietary questionnaire and had nonmissing age and blood pressure data. The upper left panel uses hexagonal binning, the upper right panel uses alpha-blending, and the two lower panels are two replicates of thinning to a subsample of 1000. The basic trend of diastolic pressure increasing until middle age and then

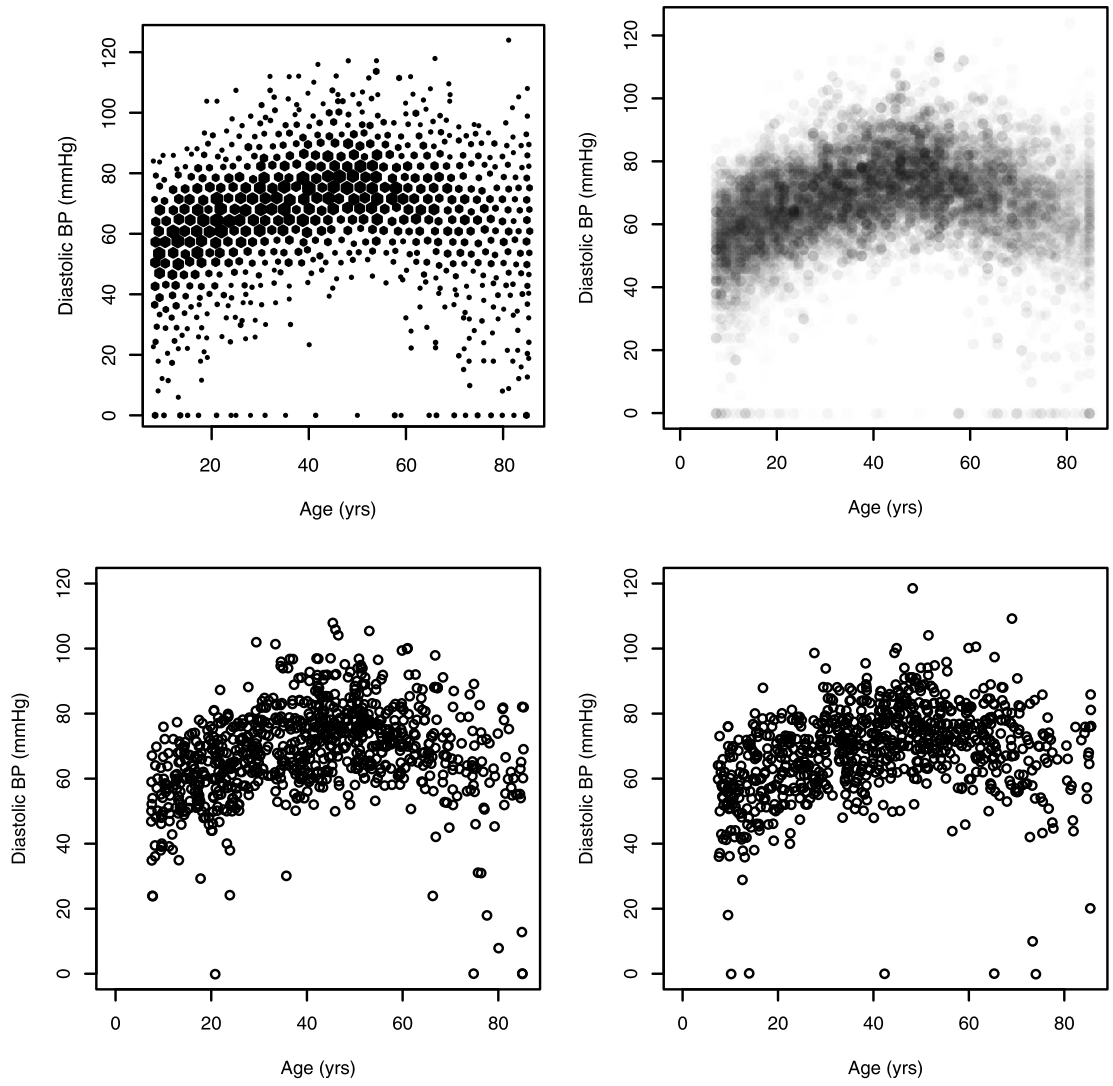

FIG. 1. *Relationship between diastolic blood pressure and age in* 13,957 *people*: *upper left panel*, *hexagonal binning*; *upper right panel*, *alpha-blending*; *lower panels*, *two replicates of thinning to a subsample of* 1000.

decreasing with stiffening blood vessels is apparent in all four graphs, as is the small group with blood pressure measured as zero—not a recording error, but a known problem with the measurement technique. The use of a minimum hex size in the left panel makes individual outliers more visible; the alpha-blending makes variations in weights more apparent. The recoding of ages over 85 to 85 is detectable in the hexagonally binned plot, and clear with alpha-blending.

Korn and Graubard (1998) also describe how to extend kernel smoothers and local regression smoothers to complex survey data: the kernel weights are multiplied by the sampling weights, so that a local mean smoother $m(x)$ for y with kernel k and bandwidth δ is

$$m(x) = \sum_{i=1}^{n} w_i w_i^K y_i,$$

$$w_i^K = \frac{k((x - x_i)/\delta)}{\sum_{i=1}^{n} k((x - x_i)/\delta)}.$$

An alternative approach to smoothing is to use regression splines, which is especially useful for quantile smoothing, as weighted quantile regression (Koenker and Basset, 1978) can then be used.

There appears to be little formal study of bandwidth selection for smoothers with sampling weights. A simple approach is to ignore the weights and use the bandwidth that would be appropriate if the data were from

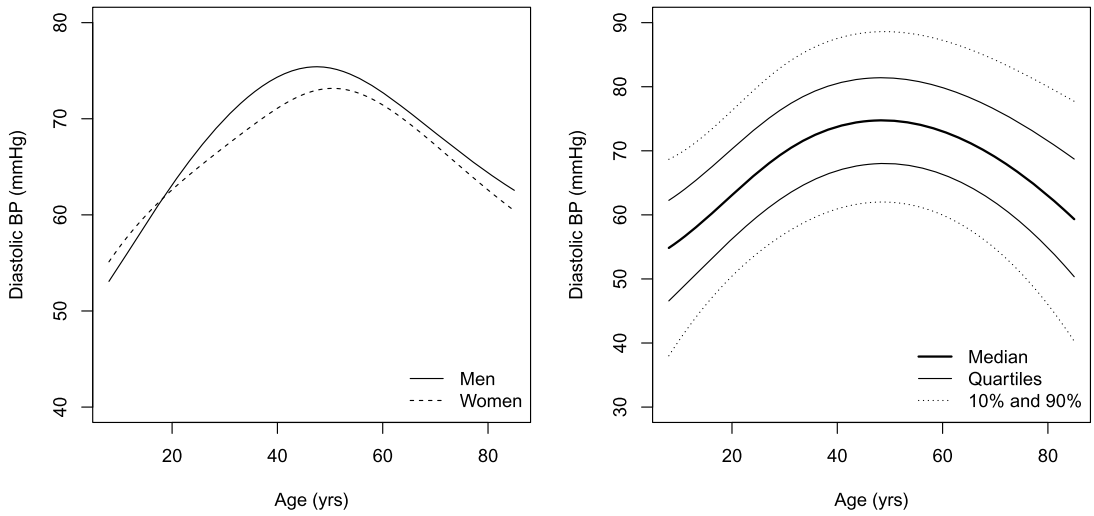

FIG. 2. *Relationship between diastolic blood pressure and age in* 13,957 *people*: *left panel*, *local-linear mean smoother separately for men and women*; *right panel*, *quantile regression splines for* 10*th*, 25*th*, 50*th*, 75*th*, 90*th percentiles*, *with* 4 *degrees of freedom*.

an i.i.d. sample; this appears to work reasonably well in practice.

Standard error estimation for these smoothers appears to be more challenging. For the local polynomial smoother, a similar approach to that under i.i.d. sampling should be possible: Harms and Duchesne (2010) give an asymptotic expression for the variance, and this could be estimated from the data. For the quantile smoother, a resampling method is likely to be needed.

Figure 2 shows the mean diastolic blood pressure by age for men and women, using a weighted local-linear regression, and the median, quartiles, and 10th and 90th percentiles using quantile regression and cubic spines with 4 equally-spaced internal knots.

The same plotting techniques can be used for residuals, partial residuals and other diagnostic model summaries involving scatterplots and smoothers.

3. PSEUDO-LIKELIHOOD ESTIMATION

3.1 Basic Weighted Estimation

With complete data on the population, we would solve the score equations:

$$\begin{aligned}\bar{U}(\beta) &= \sum_{i=1}^{N} U_i(\beta) \\ &= \sum_{i=1}^{N} \boldsymbol{x}_i \frac{1}{g'(\mu_i)V(\mu_i)}\big(y_i - \mu_i(\beta)\big) = 0,\end{aligned}$$

which are unbiased estimating equations (Godambe, 1960) for the true parameter β_0, and obtain the census parameter $\tilde{\beta}_N$. The classical design-based estimator (Fuller, 2009, Binder, 1983) solves

$$\begin{aligned}\widehat{\bar{U}}(\beta) &= \sum_{i=1}^{N} w_i R_i U_i(\beta) \\ &= \sum_{i=1}^{N} R_i w_i \boldsymbol{x}_i \frac{1}{g'(\mu_i)V(\mu_i)}\big(y_i - \mu_i(\beta)\big) = 0,\end{aligned} \tag{3.1}$$

which are unbiased estimating equations for $\tilde{\beta}_N$ if $E_\pi[w_i R_i] = 1$. Given a suitable law of large numbers and central limit theorem (Fuller, 2009), standard arguments based on smoothness can be used to show the estimator $\hat{\beta}_n$ is asymptotically normal and consistent for β_0 when the superpopulation model is correctly specified, and for β^* more generally (van der Vaart, 1998, Binder, 1983).

This is the approach underlying the regression modules in all the major statistical packages for survey analysis. It was first developed by Fuller (1975) for linear regression, and extended to more general regression models by Binder (1983). A more extensive discussion of the development can be found in Chapters 2 and 3 of Chambers and Skinner (2003).

The variance of $\hat{\beta}_n$ is the sum of two components: the finite-population sampling variance of $\hat{\beta}_n$ around $\tilde{\beta}_N$, of order n^{-1}, and the model-based sampling variance of $\tilde{\beta}_N$ around β_0, of order N^{-1}. When $n \ll N$, the latter term is often ignored; in NHANES, for example, n is less than 10^5, and N is greater than 10^8. In this setting, the variance of $\sum_{i=1}^{N} w_i R_i U_i(\beta)$ can be

computed at $\beta = \hat{\beta}_n$ by the Horvitz–Thompson variance estimator (Horvitz and Thompson, 1952). A standard delta-method argument (Binder, 1983), gives the "sandwich" form $A^{-1}BA^{-1}$ for the estimated variance of $\hat{\beta}_n$, with

$$A = \sum_{i=1}^{N} w_i R_i \frac{\partial U_i(\beta)}{\partial \beta}\bigg|_{\beta=\hat{\beta}_n}$$

and

$$B = \widehat{\text{var}}_\pi \left[\sum_{i=1}^{N} w_i R_i U_i(\beta) \right].$$

If the variance of $\tilde{\beta}_N$ is not negligible, it can be estimated from the estimated population Fisher information

$$\widehat{\text{var}}_P[\tilde{\beta}_N] = \left(\sum_{i=1}^{N} w_i R_i \frac{\partial U_i}{\partial \beta} \right)$$

and added to the finite-population sampling variance; alternatively, the middle term B of the sandwich may be replaced by a variance estimator for totals under two-phase sampling, as in Särndal, Swensson and Wretman (2003), Result 9.3.1, or Beaumont, Béliveau and Haziza (2015).

One important use for the variance is Wald tests: if β is partitioned as (β_1, β_2) then $Q = \beta_1^T \widehat{\text{var}}[\hat{\beta}_1]^{-1}\beta_1$ has an asymptotic χ^2 distribution with degrees of freedom q equal to the dimension of β_1 under the null hypothesis $\beta_1 = 0$. A better approximation is an F distribution with q numerator and m denominator degrees of freedom: $Q/q \sim F_m^q$, where m here is the so-called "design degrees of freedom", the number of primary sampling units minus the number of strata. In principle, the degrees of freedom should depend on p, but using $m - p$ as the denominator degrees of freedom tends to be conservative—it would be correct if all p covariates were constant within PSUs and only varied between PSUs. When m is small and p is moderate, $m - p$ can be very small or even negative without the distribution of Q becoming degenerate. For example, the 2003–2004 wave of NHANES has 15 strata with two primary sampling units in each stratum, giving $m = 30 - 15 = 15$, and a regression model with 15 parameters would not be exceptional.

Rust and Rao (1996) develop expressions for the denominator degrees of freedom based on the variance of the variance estimator, but further research is needed into simpler rules of thumb; these are likely to depend on the ratio of between-PSU to within-PSU variance of the predictors in the model. However, the working likelihood ratio tests described in Section 4 seem to have better operating characteristics than Wald-type tests (Thomas and Rao, 1987, Lumley and Scott, 2013). When the number of primary sampling units is small it would be attractive to have a simulation-based approach to estimating either the degrees of freedom or the whole null distribution. This is not straightforward because the distribution is sensitive to the between-cluster variation in both X and Y, and these are not well estimated with few clusters.

We can also develop an analogue to the score test. Let $\hat{\beta}_{(0)}$ be the solution to equation (3.1) under the constraint $\beta_1 = 0$. Rao, Scott and Skinner (1998) used $\widehat{\tilde{U}}(\hat{\beta}_{(0)})^T C^{-1} \widehat{\tilde{U}}(\hat{\beta}_{(0)})$, where C is an estimate of $\text{var}(\widehat{\tilde{U}}(\hat{\beta}_{(0)}))$, as the test statistic, with an asymptotic χ_p^2 distribution. With the correct choice of variance estimate, this test is invariant under transformations of the parameter, unlike the Wald test. Like the Wald test, however, the score statistics can become unstable when q is large. Following the work of Rao and Scott (1981) for loglinear models, the authors also suggested an alternative with C replaced by its equivalent under random sampling, a "plug-in" version of the score test, which is asymptotically equivalent to the working likelihood ratio tests discussed in Section 4.

When using software for generalised linear models that is not written with sampling weights in mind, the point estimates will still be the solutions to equation (3.1). The reported standard errors will be incorrect, but if the weights are scaled to sum to the sample size, the reported standard errors will typically be of the right order. Before the wide availability of software that could use sampling weights correctly, this was an important fact; over the past decade it has become much less useful.

3.2 Weights: Efficiency and Robustness

The key distinction in considering the need for weights is between endogenous and exogenous sampling schemes (DuMouchel and Duncan, 1983, Solon, Haider and Wooldridge, 2013). In an exogenous sampling scheme, R is independent of Y conditional on a set of variables X that is appropriate to include as predictors in the model; in an endogenous sampling scheme it is not. Using the econometric terminology, rather than the terms "informative" and "noninformative", emphasises that the conditioning variables must not merely be available, but must be suitable for inclusion in the model.

By the same argument as justifies propensity scores (Rosenbaum and Rubin, 1983), if the design variables are exogenous it is sufficient to condition on the weights w_i rather than X_i. The advantage of this is that the weights are *always* available; there remains the question of whether they are exogenous. In addition, conditioning on the weights rather than the design variables will complicate the interpretation of regression coefficients. In practice, this idea appears to have been used more for testing model assumptions (DuMouchel and Duncan, 1983, Wu and Fuller, 2005) or for model-based small-area inference (Verret, Rao and Hidiroglou, 2015).

If sampling is exogenous, the nonsampled fraction of the population is missing at random (Rubin, 1976). If, additionally, the mean model is correctly specified, weighted and unweighted regressions will give estimators consistent for the same parameter, so it is meaningful to compare these estimators just based on their variance. The unweighted estimator weights observations proportional to their precision, and so will be more efficient that the estimator using sampling weights. A standard rule of thumb when stratification can be ignored is that the relative efficiency of the unweighted estimator is approximately $1 + \mathrm{cv}(w)$, where $\mathrm{cv}(\cdot)$ is the coefficient of variation (Korn and Graubard, 1999, p. 173). In addition, when weighted and unweighted approaches both give (asymptotically) valid confidence intervals, the coverage will typically be closer to nominal for the unweighted estimator, partly because it has more degrees of freedom for variance estimation and partly because the sandwich variance estimator treats the weights as deterministic, when they actually depend on nonresponse and frame errors through raking or post-stratification.

However, even when the sampling is ignorable for estimating β from the marginal distribution of $Y|X$, it will often not be ignorable for estimating the variance of $\hat{\beta}$. Surveys involving in-person interviews tend to use cluster sampling, at least in areas of lower population density. The sampling indicators R_i, R_j will be correlated when i and j are in the same sampling cluster, and Y_i and Y_j are likely also to be correlated when i and j are geographically close.

When weights are not used, fitting a suitable mixed model to account for clustering will give valid standard errors for the regression coefficients. When the link function $g(\cdot)$ is linear, the mixed model parameters are the same marginal regression coefficients as in equation (1.1). For general link functions, the change of model will typically change the target of inference, but in a way that is familiar to modern statisticians and has been extensively discussed. Using sampling weights in a mixed model is much more complicated; we discuss this briefly in Section 7.

In principle, it is possible to test whether the weights are needed in the marginal model: if $\hat{\beta}$ is the weighted estimate and $\hat{\beta}_U$ the unweighted estimate, then $D = \hat{\beta} - \hat{\beta}_U$ was proposed as a test statistic by DuMouchel and Duncan (1983). If the unweighted estimator is fully efficient, their proposal is an example of the Hausman specification test (Hausman, 1978). In practice, the power of the test will be poor for contiguous alternatives, where the mean and standard deviation of D are of the same order.

Indeed, suppose $\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, \sigma^2 + \omega^2)$ and that the regression model is misspecified so that $\sqrt{n}(\hat{\beta}_U - \beta^*) \xrightarrow{d} N(\delta, \sigma^2)$. We will have $\sqrt{n}(D - \delta) \xrightarrow{d} N(0, \tau^2)$, where $\tau^2 \geq \omega^2$. If $\delta^2 = \omega^2$, the mean squared errors of $\hat{\beta}_U$ and $\hat{\beta}$ as estimators of β^* will be equal, but the test based on D will have noncentrality parameter $\delta^2/\tau^2 = \omega^2/\tau^2 \leq 1$ and will not have useful power for rejecting $\delta = 0$.

Worse, using a pre-test for the importance of weights is likely to affect the operating characteristics of subsequent hypothesis tests. There does not seem to have been systematic study of this issue in sampling, but Guggenberger (2010a, 2010b) shows in two other contexts that a two-step procedure with a Hausman specification pre-test can make the Type I error of the tests for $\beta = 0$ arbitrarily high.

While the formal test based on $\hat{\beta}_U - \hat{\beta}$ is of limited use, computing both estimates is often valuable, and may help the analyst understand how the sampling and the regression relationship under study interact.

3.2.1 *Example.* We consider two models from the same NHANES dataset used in Section 2: a logistic regression model for isolated systolic hypertension and a linear model for dietary sodium intake.

Isolated Systolic Hypertension (ISH) is defined by systolic blood pressure over 140 mmHg with diastolic blood pressure below 90; it becomes common with increasing age. We used as predictors a linear spline in age with knots at 50 and 65 years, a set of indicator variables for five-level race/ethnicity, gender and the gender:age interactions, and dietary sodium intake. We compared the sampling-weighted logistic regression model to an ordinary logistic regression model and to a generalised linear mixed model with random intercept for each primary sampling unit, fitted with the

TABLE 1

Comparison of standard errors from logistic regression models for isolated systolic hypertension (ISH), from NHANES: unweighted logistic regression with sandwich estimator, logistic-normal mixed model with random intercept for each primary sampling unit, design-weighted logistic regression

	Unweighted glm	Mixed model	Weighted
(Intercept)	1.26	1.27	1.65
Age (per extra year, in men)			
Age (≤ 50)	0.19	0.19	0.36
Age (50–65)	0.26	0.26	0.39
Age (> 65)	0.20	0.20	0.28
Female	0.82	0.83	1.20
Race:ethnicity compared to Mexican Hispanic			
Other Hispanic	0.24	0.24	0.37
Non-Hispanic Black	0.09	0.11	0.19
Non-Hispanic White	0.11	0.12	0.19
Other	0.19	0.19	0.31
Sodium	0.02	0.02	0.04
Age (per extra year, difference between women and men)			
Age (≤ 50)	0.14	0.14	0.28
Age (50–65)	0.16	0.16	0.25
Age (> 65)	0.12	0.12	0.16

TABLE 2

Comparison of linear regression models for sodium intake in grams per day, from NHANES: linear regression with sampling weights, linear regression without weights, linear mixed model with random intercept for each primary sampling unit. The national median is 3.1 g/day; the recommended daily maximum is 3 g/day

(g/day)	Weighted	Unweighted	Mixed model
(Intercept)	7.83	8.58	8.57
Age (≤ 50) per 10yrs	0.22	0.34	0.34
Age (50–65) per 10yrs	−0.68	−0.98	−0.98
Age (> 65) per 10 yrs	−0.21	−0.17	−0.17
Race:ethnicity compared to Mexican Hispanic, in men			
Other Hispanic	−0.26	−0.12	−0.08
Non-Hispanic Black	0.45	0.51	0.53
Non-Hispanic White	0.04	0.25	0.29
Other	0.13	0.24	0.26
Female	−0.75	−0.54	−0.54
Race:ethnicity compared to Mexican Hispanic, in women			
Other Hispanic	−0.13	−0.04	−0.01
Non-Hispanic Black	0.08	0.11	0.13
Non-Hispanic White	−0.01	0.15	0.19
Other	0.12	0.15	0.17

lme4 package for R (Bates et al., 2015). All of the code is in the `github` repository for the paper.

In this example, the three models gave very similar point estimates (not shown) except for the lowest age category where the slope was greater with weighting, but the weighted standard errors were large than the unweighted ones (Table 1). ISH is more common in non-Hispanic Whites than Hispanics, and less common in non-Hispanic Blacks; it increases with age, and this increase happens earlier in women than in men. There was little impact on standard error estimates of including the random intercept. In some ways, a more appropriate comparison of standard errors would use a model with random slopes for all variables, but this model did not converge.

The similarity between the models suggests that the design variables and spatial correlation in NHANES do not have an important association with hypertension conditional on age, gender and race/ethnicity. It would be reasonable to prefer the mixed model in this setting.

Higher dietary sodium intake is believed to be a risk factor for hypertension. The US median intake estimated from this NHANES sample is 3.1 g/day; the recommended daily maximum is 3 g/day. We fitted models for dietary sodium using the linear spline in age, and interactions between gender and the five race/ethnicity categories, and again we fitted a sample-weighted model, an ordinary linear model, and a linear mixed model with random intercepts for each primary sampling unit. Parameter estimates are shown in Table 2. Sodium consumption was increasing with age below 50 and decreasing with age above 50. There was a strong race:gender interaction, with non-Hispanic Black men, but not women, having higher intakes than the other race:ethnicity:gender groups by about 0.4 g/day.

The unweighted models for sodium intake had importantly different parameter estimates than the weighted models, perhaps reflecting greater regional variation in diet not captured by race:ethnicity and age, suggesting that the weighted model would be preferred.

3.2.2 *Stabilised weights.* When the mean model is correctly specified and weights depend only on x, omitting them does not affect the target of inference β_0, as we have noted earlier. More generally, multiplying or dividing the weights by any factor depending only on x also does not change β_0. The idea of dividing w_i by a function $h(x)$ chosen to minimise the variation of the resulting weights, and thus to improve efficiency, has been developed independently on at least three occasions.

Magee (1998) proposed taking a functional form $h(x) = h(x;\theta)$ and choosing θ to minimise the estimated asymptotic variance of $\hat{\beta}$, as did Skinner and Mason (2012). Robins, Hernán and Brumback (2000) proposed an estimate of $E[\pi_i|X = x]$ as $1/h(x)$. Pfeffermann and Sverchkov (1999) proposed an estimate $h(x) = E[w_i|X = x]$.

Robins, Hernán and Brumback (2000) coined the term "stabilized weights" for $w_i/h(x_i)$, which is standard in the causal-inference literature and which we recommend for use more broadly.

3.2.3 *Example.* In the logistic regression model for isolated systolic hypertension from Section 3.2.1, we computed stabilized weights, estimating $h(x)$ by a regression of weights on predictors in a Gamma generalized linear model with log link.

Figure 3 shows the estimated reduction in variance when stabilized weights are used and compares it to the estimated reduction in variance from a generalised linear mixed model with no weights. The stabilised weights always give a higher variance, but the margin differs between parameters.

3.2.4 *Calibration of weights.* It is often the case that some auxiliary variables A, or at least population totals of these variables, are available on the whole population. For example, a national census may provide the joint distribution race/ethnicity, age, sex and income band for the whole population or for smaller geographical subdivisions. *Calibration*, also called *(generalised) raking*, is an approach to using this population data (Deville and Särndal, 1992, Särndal, 2007). As a computational method, calibration is closely related to direct standardisation of rates, though direct standardisation is used to reduce bias in crude comparisons by reweighting to a common *external* standard population and calibration involves reweighting to the population from which the sample was taken.

If the sampling units are individuals, calibration is essentially the same as augmented inverse probability weighted estimation and to the technique of using estimated rather than observed weights (Robins, Rotnitzky and Zhao, 1994); the weighted empirical likelihood approach of Chaudhuri, Handcock and Rendall (2008) and the estimating equation/projection approach of Chen and Chen (2000) are very closely related. Rao, Yung and Hidiroglou (2002) appears to be the first consideration of model fitting after calibration with general sampling schemes.

In the simple case with population counts for a set of discrete categories, calibration reduces to post-stratification. That is, a scaling factor is applied to the weights for all observations in a category so that the estimated population total for the category matches the observed total. Calibration extends this idea by matching observed and estimated population totals for any set of so-called *auxiliary variables*.

In large surveys, the primary aim of calibration is reducing nonresponse bias in means and totals. It is routine for public-use data to already be calibrated to census (or administrative) totals for age groups, sex, geo-

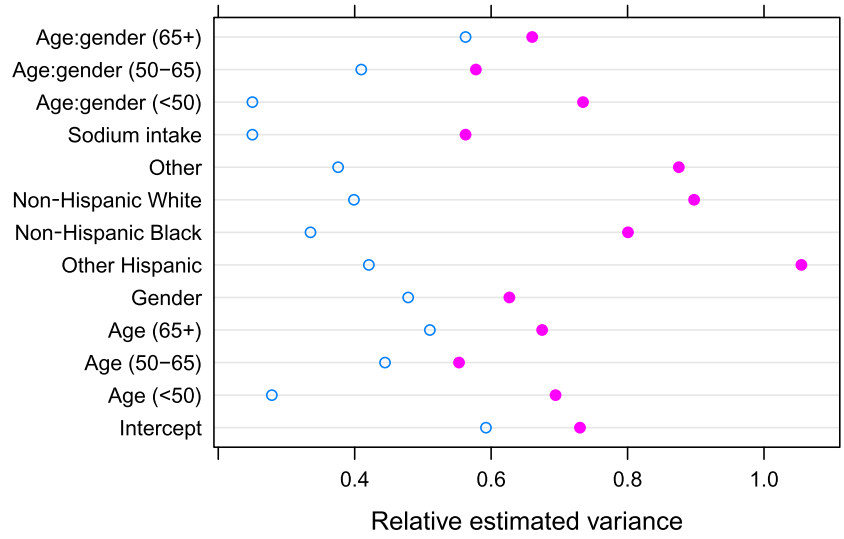

FIG. 3. *Variance reduction from stabilised weights and from not using weights, in logistic model for isolated systolic hypertension. Open circles show estimated ratio of variances for parameter estimates in a generalised linear mixed model to those in a sampling-weighted model; closed circles show the ratio for a logistic model with stabilized weights compared to the sample-weighted model.*

graphic region and in some countries, race or ethnicity. When estimating a population total, calibration reduces nonresponse bias to the extent that the auxiliary variable explains the correlation between nonresponse and the variable of interest. Calibration also increases precision to the extent that the auxiliary variable is correlated with the variable of interest; it can be viewed as a form of imputation.

Calibration tends to be of less benefit when fitting regression models. As Lumley, Shaw and Dai (2011) shows, it is helpful to think of $\hat{\beta} - \beta^*$ as approximately an estimated population total: the total of the influence functions. Influence functions (closely related to the $\Delta\beta$ deletion diagnostics) can be defined by

$$\sqrt{n}(\hat{\beta} - \beta) = \sum_{i=1}^{n} w_i IF_\beta(y, x) + o_p(1). \tag{3.2}$$

The influence function is a linear approximation to the change in $\hat{\beta}$ when an observation at (x, y) is deleted (van der Vaart, 1998, Chapter 20). In the context of generalised linear models, the influence functions are given by $A^{-1}U_i$ as defined in Section 3.1 and so are a linear function of $x_i(y_i - \mu_i(\beta))$. Most estimators of traditional statistical interest can be written in the form of equation (3.2), but some modern sparse estimators, such as the lasso, are not regular or asymptotically linear and have no influence-function decomposition.

Calibration is helpful only when the auxiliary variables are correlated with the variable being summed; in this case, the influence function. The influence functions are nearly uncorrelated with Y and X, so these variables or surrogates for them will not be good auxiliary variables; we need auxiliary variables that are linearly correlated with the influence function for the regression parameter of interest (Breslow et al., 2009a, Lumley, Shaw and Dai, 2011). The main setting when this is plausible is subsampling from an existing cohort, whether from a research study, an insurance system, or national adminstrative records. Breslow et al. (2009b) describes one strategy for achieving useful correlation when rich data are available on the population (or cohort) from which the data are sampled. Støer and Samuelsen (2012) and Rivera and Lumley (2015) describe similar approaches for other designs.

Realising the gains in precision from calibration requires standard error estimators that take these precision gains into account. These are now readily available for generalised linear models and the Cox model, but for new models it may be easiest to work via resampling. In survey statistics, resampling estimators analogous to the jackknife and bootstrap are typically written in terms of sets of resampling weights or *replicate weights*. For example, in a cluster jackknife sample, the weight for observations in one cluster will be set to zero and the weights for observations in other clusters increased to compensate. If each set of replicate weights is calibrated to the same population totals as the sample weights, the resampling standard errors will be correct (Valliant, 1993).

Calibration improves asymptotic efficiency for the same target parameter. It makes essentially no modelling assumptions, and the asymptotic efficiency of the calibrated estimator is never worse than that of the uncalibrated estimator. In comparison, stabilising weights can lead to further improvements in efficiency, but can introduce bias unless the mean model for Y is correctly specified, and a sufficiently poor choice of $h(\cdot)$ can lead to a variance increase.

Calibration weakens the missing at random assumption on nonresponse by allowing dependence on (possibly endogenous) auxiliary variables. Even so, the assumption is likely to be untrue. Kott and Chang (2010) and Pfeffermann and Sikov (2011) discuss two ways to go further, which unavoidably require untestable model assumptions.

4. WORKING LIKELIHOOD TESTS AND INFORMATION CRITERIA

Although there is no natural and straightforward likelihood function for survey data, it is possible to construct an analogue of the likelihood-ratio test based on the pseudo-likelihood,

$$\begin{aligned}\widehat{\ell}(\beta) &= \sum_{i=1}^{N} w_i R_i \log f(y|x;\beta) \\ &= \sum_{i=1}^{N} w_i R_i \ell_i(\beta), \quad \text{say},\end{aligned}$$

that has many of the same properties. This extends the work of Rao and Scott (1981) for loglinear models under complex sampling and Rotnitzsky and Jewell (1990) for generalised linear models with clustering.

Suppose that we are interested in testing the hypothesis $H_0 : \beta_1 = 0$, where β is partitioned as (β_1, β_2), as in the section on Wald Tests in Section 3.1. Then our working likelihood ratio test statistic is given by

$$\Lambda = 2\{\widehat{\ell}(\widehat{\beta}) - \widehat{\ell}(\widehat{\beta}_{(0)})\}, \tag{4.1}$$

where $\widehat{\beta}_{(0)}$ is the solution to equation (3.1) under the constraint $\beta_1 = 0$.

Lumley and Scott (2014) showed that, under the regularity conditions of Theorem 1.3.9 in Fuller (2009),

$\Lambda \sim \sum_1^q \delta_i Z_i^2$ asymptotically under H_0, where $Z_1, \ldots, Z_q$ are independent standard normal random variables and and $\delta_1, \ldots, \delta_q$ are the eigenvalues of $\Delta = (\mathcal{I}^*_{11} - \mathcal{I}^*_{12}\mathcal{I}^{*-1}_{22}\mathcal{I}^*_{21})V_1^*$. Here, $V_1^* = V_1(\beta^*)$ is the asymptotic covariance matrix of $\sqrt{n}\{\widehat{\beta}_1 - \beta_1^*\}$ and $\mathcal{I}^* = \left(\begin{smallmatrix}\mathcal{I}_{11} & \mathcal{I}_{12}\\ \mathcal{I}_{21} & \mathcal{I}_{22}\end{smallmatrix}\right) = \mathcal{I}(\beta^*)$. The argument uses a second-order Taylor series approximation; the linear term vanishes at the maximum, and the highest-order remaining term is a quadratic form in asymptotically-normal variables.

If the sample had been a random sample from the superpopulation, then $\mathrm{Var}(\widehat{\beta})$ would be equal to $\mathcal{I}^{-1}$. Using the standard form for the inverse of a partitioned matrix, it follows that $\mathrm{Var}(\widehat{\beta}_1)$ would be equal to

$$\left(\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1} = V_{01},$$

say. Thus, we can write the matrix Δ in the form $\Delta = V_{01}^{*\,-1}V_1^*$. By analogy with the simple scalar case, we call Δ the "design-effect matrix" and the eigenvalues, $\delta_1, \ldots, \delta_q$, "generalised design effects", as in Rao and Scott (1984, 1981).

The value of Λ is very sensitive to the scaling of the weights. Rao and Scott (1981) suggested dividing Λ by the average eigenvalue, $\bar{\delta}$, to get around this and we recommend this form for display in the output of a regression program. If the eigenvalues are all equal, as in some special designs, the asymptotic null distribution of $\Lambda/\bar{\delta}$ is exactly χ^2_q. Otherwise, it has the correct mean, q, but the associated p-value will need to be corrected.

A Satterthwaite approximation to the distribution of Λ is standard and surprisingly accurate: $\Lambda/\bar{\delta} \sim \chi^2_\nu$ with $\nu = (\sum_i \delta_i^2)/(\sum_i \delta_i)^2$. When higher accuracy is required, for example, for large-scale multiple testing in genomics, options with accurate free-software implementations include an infinite series (Farebrother, 1984), a method based on characteristic functions (Davies, 1980), and a saddlepoint approximation (Kuonen, 1999).

Under the null hypothesis that $\beta_1^* = 0$, 2Λ would have expectation $2q$ under i.i.d. sampling. Under complex sampling, it has expectation $2q\bar{\delta}$, motivating

$$\mathrm{dAIC} = 2\Lambda - 2p\hat{\bar{\delta}},$$

where $\hat{\bar{\delta}} = \mathrm{tr}(\mathcal{I}V)/p$ is the average design effect for the full model, as a design-based version of AIC. Lumley and Scott (2015) show that minimising *dAIC* minimizes the expected Kullbeck–Leibler divergence from the true model and has connections to minimising cross-validated prediction error in the same way that minimising AIC does.

TABLE 3
dAIC and dBIC in five models for isolated systolic hypertension fitted to the NHANES data

	dAIC	dBIC
Age	7786	7756
+ race/ethnicity	7772	7750
+ gender	7764	7729
+ gender:age	**7664**	**7727**
+ sodium	7670	7734

An analogue to BIC is less straightforward, as BIC is derived from a Laplace approximation to posterior probabilities and relies on the full log-likelihood. Simply changing the $2p\hat{\bar{\delta}}$ penalty to $p\hat{\bar{\delta}}\log n$ does not preserve the derivation of BIC.

Fabrizi and Lahiri (2007) constructed a penalised Wald statistic and showed it was asymptotically equivalent to BIC based on the full log-likelihood if the design-based estimator is asymptotically efficient. Lumley and Scott (2015) derived essentially the same criterion without the assumption of efficiency, from a coarsened-Bayesian argument. They reduced the data to the parameter estimates $\hat{\beta}$ and used the asymptotic Gaussian likelihood for these estimates under each model to construct posterior probabilities and the BIC.

4.1 Example

In both of the models for isolated systolic hypertension described in Section 3.2.1 and the models for sodium intake in Table 2, there is good qualitative agreement between Wald tests and working likelihood ratio tests, with the design effect $\hat{\bar{\delta}}$ being between 2 and 3 for the ISH models and between 5 and 6 for the sodium-intake models. Table 3 shows dAIC and dBIC for the ISH models.

The two criteria agree on the best model, but as would be expected dBIC is more supportive of the simpler model without gender:age interactions than dAIC is, and gives it second place.

5. OTHER WAYS TO USE THE WEIGHTS

Under the idealised model of probability sampling without nonresponse, the sampling probabilities π_i contain all the information about the marginal relationship between R_i and Y_i, and even in a more realistic setting, the weights w_i after adjustment for

nonresponse and frame errors contain all the readily-available information. Modelling the weights is likely to be useful in model-based inference.

Kim and Skinner (2013) built on work outside the regression context by Beaumont (2008), Pfeffermann and Sverchkov (1999), and Pfeffermann, Krieger and Rinott (1998), to construct "smoothed weights". In addition to the observation that an arbitrary multiplicative function of x can be introduced, they note that variation in the weights that is independent of Y given x is uninformative. They propose estimating $\tilde{d}_j i = E[w_i|y_i, x_i, R_1 = 1]$ to capture the informative component of the weights. If the regression model and the model for $\tilde{d}_i$ are both correctly specified, w_i can be replaced by $\tilde{d}_i h(x_j)$ for an arbitrary function $h(\cdot)$, which can then be chosen to minimise the variance of $\hat{\beta}$ as in Section 3.2. Because the limiting value of $\hat{\beta}$ depends on correct specification of the model for $\tilde{d}_i$, estimation of any parameters in that model contributes to the variance of $\hat{\beta}$; Kim and Skinner (2013) describe how to estimate the variance.

Elliott (2007) and Elliott (2009) describe a Bayesian approach to generalized linear models that relies on the sampling weights to provide information about informative sampling, but uses model-based shrinkage and Winsorisation to reduce the variability in the weights in order to increase precision.

6. MAXIMUM LIKELIHOOD ESTIMATION

The full likelihood for a regression model under complex sampling involves the joint distribution of R and Y and is typically intractable, but maximum likelihood estimation is available in a few important scenarios. In addition to being of practical use, these scenarios allow direct comparisons of likelihood estimators with weighted estimators, on an equal footing.

Case-control sampling is the oldest and most important design where maximum likelihood estimation is available. In a standard case-control design, sampling is stratified on a rare binary outcome Y ascertainable for everyone in a population. The sample includes everyone with $Y = 1$ (cases) and a small fraction π_0 of those with $Y = 0$ (controls) so that the number of controls is a small multiple (usually 1–5) of the number of cases. Under a model with an arbitrary marginal distribution of predictor variables x and with Y satisfying a logistic regression model

$$\text{logit}\, P[Y = 1|X = x] = x\beta,$$

the semiparametric maximum likelihood estimator $\hat{\beta}_{\text{mle}}$ is obtained by unweighted logistic regression (Prentice and Pyke, 1979), and this estimator is semiparametric-efficient (Breslow, Robins and Wellner, 2000). Case-control sampling is the extreme case of endogenous sampling: the only design variable that affects sampling weights is Y, and it would not make sense to include Y as a predictor in regression models.

In a case-control study the coefficient of variation of the weights will always be high (close to 1 when the cases are rare and the numbers of cases and controls in the sample are equal), but the coefficient of variation within sampling strata is zero. General heuristics are not known for the relative efficiency of the weighted estimator, which can be as high as 100% or can be arbitrarily low. Full efficiency occurs when the model is saturated or when all parameters apart from the intercept are zero; the easiest way to achieve low efficiency is where there is a very strong effect of a continuous predictor.

For example, Breslow and Day (1980) give two data sets from a study of alcohol and tobacco as risk factors for esophageal cancer in Ille-et-Vilaine, France: one with continuous exposures and one with grouped exposures. Reanalysing the data with weights estimated from the size of the contemporary population makes little or no difference to the grouped-exposure models but leads to increases of 50% in estimated variance for the alcohol and tobacco coefficients in the continuous-data model. Figure 4 shows the variance ratio for the model with main effects of age, alcohol and tobacco.

The maximum likelihood estimator is not design-consistent: if the model is misspecified, $\hat{\beta}_{\text{mle}}$ does not converge to β^*. Maximum likelihood estimation for the logistic model under case-control sampling is universal in epidemiology and biostatistics, but is somewhat controversial elsewhere, for this reason. Scott and Wild (2002) summarise this debate and discuss the interpretation of $\hat{\beta}_{\text{mle}}$ under model misspecification.

When Y is the only variable available for the whole population, the weighted estimator is already the efficient design-consistent estimator and calibration thus provides no gain in efficiency. Stabilized weights do provide some efficiency increase, but not (typically) to the level of the MLE, and weight smoothing provides no further improvement. Since these weight adjustments do not give a fully efficient estimator in the case-control design, they presumably do not give fully efficient estimators in other designs where the computations are less tractable.

7. DISCUSSION

"Statistical methods *need* software", as Brian Ripley has emphasized, and for many years survey analysis

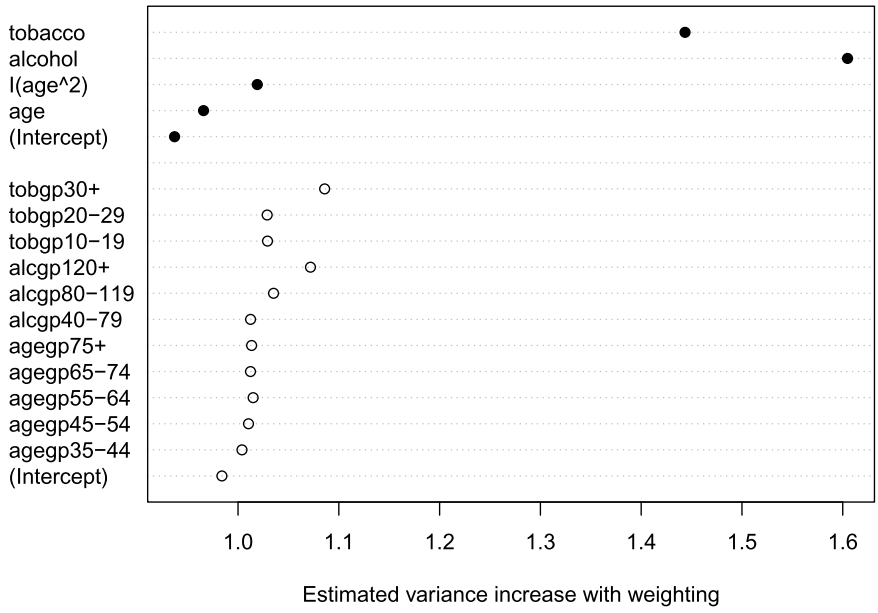

FIG. 4. *Estimated variance increase from weighting, for data from a case–control study of esophageal cancer: open circles, grouped exposure; filled circles, continuous exposure.*

used specialised software running on high-end hardware. There has been a lot of progress over the past decade, however, and all general purpose packages can now do estimation and Wald tests at least for linear and logistic models.

Alan Zaslavsky maintains a list of survey data analysis software at http://www.hcp.med.harvard.edu/statistics/survey-soft/. Among these, the widest range of models appears to be available in Stata (StataCorp, 2015) and the widest range of designs in the survey package for R (Lumley, 2015). Heeringa, West and Berglund (2010) provide worked examples and discussion across multiple packages. The Rao–Scott tests for generalised linear models and the related information criteria are currently only available in the R survey package. We are not aware of any packages that currently integrate stabilised or smoothed weights into the data analysis workflow.

All the design-based inference we discuss is for marginal models. Design-based inference even for linear mixed models is substantially more complicated. In the special case where the clusters in the model are the same as the clusters in the sampling design, there is literature on consistent estimation (Pfeffermann et al., 1998, Rabe-Hesketh and Skrondal, 2006, Rao, Verret and Hidiroglou, 2014) and some implementations (Muthén and Muthén, 2012, StataCorp, 2015). The more general case—for example, a mixed model with covariance describing kinship in a design sampled by household and administrative unit—has received little consideration and simplified approaches have been used (Lin et al., 2014, Morrison et al., 2016).

It is undoubtedly true that the simple weighted analyses we have focused on will fail to be fully efficient. They will also fail to regularize estimates for small subgroups optimally. An important area of research and development would be to make more efficient analyses of data with potentially-endogenous sampling weights routinely usable by the nonexpert. Gelman (2007) and Little (2012) discuss some of the principles and issues.

ACKNOWLEDGMENTS

Partly supported by a grant from the Marsden Fund of the Royal Society of New Zealand.

REFERENCES

BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.

BEAUMONT, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika* **95** 539–553. MR2443174

BEAUMONT, J.-F., BÉLIVEAU, A. and HAZIZA, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology* **3** 524–542.

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **51** 279–292. MR0731144

BRESLOW, N. E. and DAY, N. E. (1980). *Statistical Methods in Cancer Research*, *Volume I—The Analysis of Case-Control Studies*. IARC Publications, Paris.

BRESLOW, N. E., ROBINS, J. M. and WELLNER, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6** 447–455. MR1762555

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009a). Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences* **1** 32–49.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009b). Using the whole cohort in the analysis of case-cohort data. *Am. J. Epidemiol.* **169** 1398–1405.

CARR, D. B., LITTLEFIELD, R. J., NICHOLSON, W. L. and LITTLEFIELD, J. S. (1987). Scatterplot matrix techniques for large N. *J. Amer. Statist. Assoc.* **82** 424–436. MR0898351

CHAMBERS, R. L. and SKINNER, C. J., eds. (2003). *Analysis of Survey Data*. Wiley, Chichester. MR1978840

CHAUDHURI, S., HANDCOCK, M. S. and RENDALL, M. S. (2008). Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 311–328. MR2424755

CHEN, Y.-H. and CHEN, H. (2000). A unified approach to regression analysis under double-sampling designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 449–460. MR1772408

DAVIES, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **29** 323–333.

DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

DUMOUCHEL, W. H. and DUNCAN, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *J. Amer. Statist. Assoc.* **78** 535–543.

ELLIOTT, M. R. (2007). Bayesian weight trimming for generalized linear regression models. *Surv. Methodol.* **33** 23–34.

ELLIOTT, M. R. (2009). Model averaging methods for weight trimming in generalized linear regression models. *J. Off. Stat.* **25** 1–20.

FABRIZI, E. and LAHIRI, P. (2007). A design-based approximation to the BIC in finite population sampling. Technical Report 4, Dipartimento di Matematica, Statistica, Informatica e Applicazioni, Università degli Studi di Bergamo.

FAREBROTHER, R. W. (1984). Algorithm AS 204: The distribution of a positive linear combination of χ^2 random variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **33** 332–339.

NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and Operation of the Third National Health and Nutrition Examination Survey, 1976–1980. Number 32 in Series 1: Programs and Collection Procedures.

FULLER, W. A. (1975). Regression analysis for sample survey. *Sankhyā, Series C* **37** 117–132.

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

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951

GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.* **31** 1208–1211. MR0123385

GUGGENBERGER, P. (2010a). The impact of a Hausman pretest on the size of a hypothesis test: The panel data case. *J. Econometrics* **156** 337–343.

GUGGENBERGER, P. (2010b). The impact of a Hausman pretest on the asymptotic size of a hypothesis test. *Econometric Theory* **26** 369–382.

HARMS, T. and DUCHESNE, P. (2010). On kernel nonparametric regression designed for complex survey data. *Metrika* **72** 111–138.

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271.

HEERINGA, S., WEST, B. T. and BERGLUND, P. A. (2010). *Applied Survey Data Analysis*. CRC Press, Boca Raton, FL.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

KIM, J. K. and SKINNER, C. J. (2013). Weighting in survey analysis under informative sampling. *Biometrika* **100** 385–398. MR3068441

KOENKER, R. and BASSET, G. (1978). Regression quantiles. *Econometrica* **46** 33–50.

KORN, E. L. and GRAUBARD, B. I. (1998). Scatterplots with survey data. *Amer. Statist.* **52** 58–69.

KORN, E. L. and GRAUBARD, B. I. (1999). *Analysis of Health Surveys*. Wiley, New York.

KOTT, P. S. and CHANG, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *J. Amer. Statist. Assoc.* **105** 1265–1275. MR2752620

KUONEN, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86** 929–935.

LIN, D., TAO, R., KALSBEEK, W., ZENG, D., GONZALEZ, F. II, FERNANDEZ-RHODES, L., GRAFF, M., KOCH, G. G., NORTH, K. and HEISS, G. (2014). Genetic association analysis under complex survey sampling: The Hispanic Community Health Study/Study of Latinos. *American Journal of Human Genetics* **95** 675–688.

LITTLE, R. J. A. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics. *J. Off. Stat.* **28** 309–372.

LUMLEY, T. (2010). *Complex Surveys*: *A Guide to Analysis Using R*. Wiley, Hoboken, NJ.

LUMLEY, T. (2015). survey: Analysis of complex survey samples. R package version 3.30-3. Available at https://cran.r-project.org/package=survey.

LUMLEY, T. and SCOTT, A. J. (2013). Partial likelihood ratio tests for the Cox model under complex sampling. *Stat. Med.* **32** 110–123. MR3017887

LUMLEY, T. and SCOTT, A. J. (2014). Tests for regression models fitted to survey data. *Aust. N. Z. J. Stat.* **56** 1–14. MR3200288

LUMLEY, T. and SCOTT, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology* **3** 1–18.

LUMLEY, T., SHAW, P. A. and DAI, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *Int. Stat. Rev.* **79** 200–220.

MAGEE, L. (1998). Improving survey-weighted least squares regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 115–126. MR1625652

MORRISON, J., LAURIE, C., MARAZITA, M., SANDERS, S., OFFENBACHER, S., SALAZAR, C., CONOMOS, M., THORNTON, T., JAIN, D., LAURIE, C., KERR, K., PAPANICO-

LAOU, G., TAYLOR, K., KASTE, L., BECK, J. and SHAFFER, J. (2016). Genome-wide association study of dental caries in the Hispanic Communities Health Study/Study of Latinos (HCHS/SOL). *Human Molecular Genetics* **25** 807–816.

MUTHÉN, L. K. and MUTHÉN, B. O. (2012). *Mplus User's Guide*, 7th ed. Muthén & Muthén, Los Angeles, CA.

PFEFFERMANN, D., KRIEGER, A. M. and RINOTT, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statist. Sinica* **8** 1087–1114. MR1666233

PFEFFERMANN, D. and SIKOV, A. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *J. Off. Stat.* **27** 181–209.

PFEFFERMANN, D. and SVERCHKOV, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā, Series B* **61** 166–186. MR1720710

PFEFFERMANN, D., SKINNER, C. J., HOLMES, D. J., GOLDSTEIN, H. and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 23–40. MR1625668

PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411. MR0556730

RABE-HESKETH, S. and SKRONDAL, A. (2006). Multilevel modelling of complex survey data. *J. Roy. Statist. Soc. Ser. A* **169** 805–827. MR2291345

RAO, J. N. K. and SCOTT, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *J. Amer. Statist. Assoc.* **76** 221–230. MR0624328

RAO, J. N. K. and SCOTT, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Ann. Statist.* **12** 46–60. MR0733498

RAO, J. N. K., SCOTT, A. J. and SKINNER, C. J. (1998). Quasi-score tests with survey data. *Statist. Sinica* **8** 1059–1070.

RAO, J. N. K., VERRET, F. and HIDIROGLOU, M. A. (2014). A weighted composite likelihood approach to inference for two-level models from survey data. *Surv. Methodol.* **39** 263–282.

RAO, J. N. K., YUNG, W. and HIDIROGLOU, M. A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā, Series A* **64** 364–378. MR1981764

RIVERA, C. and LUMLEY, T. (2015). Using the whole cohort in the analysis of countermatched samples. *Biometrics* **72** 382–391.

ROBINS, J. M., HERNÁN, M. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemology. *Epidemiology* **11** 550–560.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROSENBAUM, P. L. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROTNITZKY, A. and JEWELL, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77** 485–497. MR1087838

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUST, K. F. and RAO, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Stat. Methods Med. Res.* **5** 283–310.

SÄRNDAL, C.-E. (2007). The calibration approach in survey theory and practice. *Surv. Methodol.* **33** 99–119.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (2003). *Model Assisted Survey Sampling*. Springer, Berlin. MR1140409

SCOTT, A. and WILD, C. (2002). On the robustness of weighted methods for fitting models to case-control data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 207–219. MR1904701

SKINNER, C. and MASON, B. (2012). Weighting in the regression analysis of survey data with a cross-national application. *Canad. J. Statist.* **40** 697–711. MR2998857

SOLON, G., HAIDER, S. J. and WOOLDRIDGE, J. (2013). What are we weighting for? Working Paper 18859, National Bureau of Economic Research, Cambridge, MA.

STATACORP (2015). Stata Statistical Software: Release 14. StataCorp LP, College Station, TX.

STØER, N. C. and SAMUELSEN, S. O. (2012). Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Anal.* **18** 261–283. MR2945403

THOMAS, D. R. and RAO, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *J. Amer. Statist. Assoc.* **82** 630–636. MR0898369

UNWIN, A., THEUS, M. and HOFMANN, H., eds. (2007). *Graphics of Large Datasets*: *Visualizing a Million*. Springer, New York.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *J. Amer. Statist. Assoc.* **88** 89–96. MR1212481

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

VERRET, F., RAO, J. and HIDIROGLOU, M. A. (2015). Model-based small area estimation under informative sampling. *Surv. Methodol.* **41** 333–347.

WU, Y. and FULLER, W. A. (2005). Preliminary testing procedures for regression with survey samples. In *Proceedings of the Section on Survey Research Methods* 3683–3688. Amer. Statist. Assoc., Alexandria, VA.