Fitting Regression Models to Complex Survey Data—Gelman's Estimator Revisited

Feder, Moshe
Southampton Statistical Sciences Research Institute
Building 39
University of Southampton
Southampton SO17 1BJ, U.K.
E-mail: m.feder@soton.ac.uk, moshe.feder@gmail.com

Survey Data

Survey data typically have certain characteristics which must be accounted for in their analysis. These characteristics include unequal selection probabilities, clustering, and missing data (including non-response). In this paper, we focus only on unequal selection probabilities and assume full response. For a more complete discussion see, for example, Pfeffermann & Sverchkov (2009).

Consider a model of interest with a response (dependent) variable y and a vector of explanatory (independent) variables $\boldsymbol{x}$, and let y_i and $\boldsymbol{x}_i$ be the values associated with unit i. The sample conditional distribution $f_s(y_i|\boldsymbol{x}_i)$, defined in (1) below is usually different from the population conditional distribution $f_p(y_i|\boldsymbol{x}_i)$.

Indeed, by Bayes Rule,

$$(1) \quad f_s(y_i|x_i) \overset{\text{def}}{=} f_p(y_i|x_i, i \in s) = \frac{\Pr(i \in s|y_i, x_i)}{\Pr(i \in s|x_i)} f_p(y_i|x_i).$$

Consequently, unless $\Pr(i \in s|y_i, \boldsymbol{x}_i) = \Pr(i \in s|\boldsymbol{x}_i)$ for all i, the sample and the population distributions are different: $f_s(y_i|\boldsymbol{x}_i) \neq f_p(y_i|\boldsymbol{x}_i)$. Therefore, a possible approach is to augment the set of explanatory variables by variables that determine the sample design ('design variables'), which we'll denote by $\boldsymbol{z}$. If $\Pr(i \in s|y_i, \boldsymbol{x}_i, \boldsymbol{z}_i) = \Pr(i \in s|\boldsymbol{x}_i, \boldsymbol{z}_i)$, modelling y_i on $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ mitigates the effect of the sampling design on the distribution. That leaves, however, additional explanatory variables in the model that may pose difficulty in their interpretation. Assuming the conditional distribution of $f(\boldsymbol{z}_i|\boldsymbol{x}_i)$ is known, one may integrate out $\boldsymbol{z}_i$ from the model:

$$(2) \quad f_p(y_i|\boldsymbol{x}_i) = \int f_p(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i) f_p(\boldsymbol{z}_i|\boldsymbol{x}_i)\, d\boldsymbol{z}_i,$$

an approach proposed in a recent paper by Gelman (2007). See also an earlier paper by Skinner (1994) where the special case of $\boldsymbol{z}_i = w_i$, the sampling weights was treated. (There have been a few variants of this approach in the literature—see Pfeffermann (forthcoming) for a discussion.)

In the case of linear regression, we first fit a model

$$(3) \quad E(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{x}_i + \boldsymbol{\beta}_2'\boldsymbol{z}_i + \boldsymbol{\beta}_3'\boldsymbol{r}_i,$$

where $\boldsymbol{r}_i$ is a vector of $\boldsymbol{x} \cdot \boldsymbol{z}$ interactions. Then, taking expectation conditional on $\boldsymbol{x}_i$ alone, we obtain

$$(4) \quad E(y_i|\boldsymbol{x}_i) = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{x}_i + \boldsymbol{\beta}_2' E(\boldsymbol{z}_i|\boldsymbol{x}_i) + \boldsymbol{\beta}_3'\boldsymbol{x}_i \cdot E(\boldsymbol{z}_i|\boldsymbol{x}_i).$$

In the last stage, an estimate of $E(\boldsymbol{z}|\boldsymbol{x})$ is needed. To do that, Gelman (2007) first estimates $f_p(\boldsymbol{x}_i|\boldsymbol{z}_i)$ from the sample data, then assuming known $f_p(\boldsymbol{z}_i)$ he obtains $E(\boldsymbol{z}_i|\boldsymbol{x}_i)$ by application of Bayes' Rule.

REMARK 1: For the sake of simplicity, we limit the discussion below to joint distributions unit-level variables. For example, under a more realistic scenario, the inclusion probability of unit i may depend on other units' values of $\boldsymbol{z}$. A more general discussion requires replacing the subscripts i by s where, for example $\boldsymbol{z}_s$, or even $\boldsymbol{z}_u$ (sample or population values) replaces $\boldsymbol{z}_i$.

REMARK 2: Rubin (1985) suggests that the sampling weights w_i may be used in place of the $\boldsymbol{z}$ variables when w_i are an *adequate summary* of $\boldsymbol{z}$. See Rubin (1985) for definition and details.

In the case where the objective is to fit a linear model to the response variable, applying (2) to the 'intermediate' model (3) may be unsatisfactory since the resultant model will in general, be non-linear. Furthermore, the integration (2) may not be straightforward.

Proposed Approach

We consider two scenarios: (1) Known population values $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i \in U}$ and (2) $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$ are only observed for the sample units.

(1) Known $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i \in U}$
This assumption is realistic when these values are part of census data or administrative records. First, we fit the intermediate model (3). This allows us to predict a value $\tilde{y}_i$, say, for each unit $i \in U$ in the population. For sampled units $i \in s$, $\tilde{y}_i = y_i$ is the observed value. For the non-sampled units $i \notin s$, we have the following options, depending on the purpose of the model-fitting:

1. Predict $\tilde{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1' \boldsymbol{x}_i + \hat{\boldsymbol{\beta}}_2' \boldsymbol{z}_i + \hat{\boldsymbol{\beta}}_3' \boldsymbol{r}_i$
2. Predict $\tilde{y}_i$ as above with added noise drawn from a normal distribution $N(0, \hat{\sigma}^2)$ where $\hat{\sigma}^2$ is the estimate of the variance of the residual error when regressing y on $(\boldsymbol{x}, \boldsymbol{z})$.
3. Similar to Option 2 above, but with a residual drawn with replacement from estimated residuals amongst sample units with similar covariates.

We now can fit a linear regression model to the data $\{\tilde{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\}_{i \in U}$ with $\tilde{y}_i$ as the dependent variable. Clearly Option 1 above yields more precise estimates than Options 2 & 3. However, Options 2 and 3 allow prediction of the response variable for non-sampled units. In contrast, predictions $\tilde{y}_i$ using Option 1 will be deterministically defined by $(\boldsymbol{x}_i, \boldsymbol{z}_i)$.

(2) Unknown $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i \in U}$
In the case where the values $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i \in U}$ are unknown for non-sampled units, we need to obtain predictions $\{\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{z}}_i\}_{i \in U}$ for them. This may be carried out by sampling $N - n$ units with replacement from the sample values, with probabilities proportional to $(w_i - 1)/(\sum_j (w_j - 1)$ on each draw, where w_i is the sampling weight (inverse of the selection probability) of unit i (see Pfeffermann and Sikov, 2011 for justification). We now have an imputed population $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i \in S} \cup \{\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{z}}_i\}_{i \notin S}$. Next, we proceed as in the known covariates and design data above. Predict $\tilde{y}_i$ for the non-sampled units, to obtain $\{y, \boldsymbol{x}_i, \boldsymbol{z}_i\}_{i \in S} \cup \{\tilde{y}, \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{z}}_i\}_{i \notin S}$. Finally, we fit a linear regression of y on $\boldsymbol{x}$ to obtain our estimated regression coefficients. Put in another way, the proposed estimate is $(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\boldsymbol{y}}$, where the rows in $\tilde{X}$ and the componets in $\tilde{\boldsymbol{y}}$ corresponding to sampled units are thosed of the observed units, and those corresponding to unobserved units are those imputed.

Variance Computation

We consider the case where the finite population values of $\boldsymbol{x}$ and $\boldsymbol{z}$ are known. let X_u be the matrix whose rows are the *population* units' x variables, including 1 for the intercept and let R_u be the matrix whose rows are the finite population values of $(\boldsymbol{x}, \boldsymbol{z})$ and their interaction terms. Let X_s and R_s be the rows in X_u and R_u corresponding to the sample units, and R_c (X_c) the rows in R_u (respectively, X_u) corresponding to the non-sampled units. Denote by $\boldsymbol{y}_s = (y_1, \ldots, y_n)'$ the vector of the n observed values of the response variable.
Then the proposed estimate is

$$\begin{aligned}\hat{\beta} &= (X_u'X_u)^{-1}X_u' \begin{pmatrix} \boldsymbol{y}_s \\ R_c\tilde{\gamma} + \varepsilon_c \end{pmatrix} = (X_u'X_u)^{-1}(X_s', X_c') \begin{pmatrix} \boldsymbol{y}_s \\ R_c\tilde{\gamma} + \varepsilon_c \end{pmatrix} \\ &= (X_u'X_u)^{-1}X_s'\boldsymbol{y}_s + (X_u'X_u)^{-1}X_c'R_c\tilde{\gamma} + (X_u'X_u)^{-1}X_c'\varepsilon_c\end{aligned}$$

where $\tilde{\gamma}$ is the estimated regression coefficient of the intermediate model (3), and where ε_c is a vector of added random noise (could be zero, if no noise is added). From a model-based point of view, the first term of the right-hand side is fixed. The second and third terms are of the form $A\tilde{\gamma}$ and $B\varepsilon_c$ for some A, B, are independent. Estimation of the variance of $\tilde{\gamma}$ is straightforward, and the variance of the added noise is also known.
From a randomization point of view the first term is random. Note that we can write $X_s'\boldsymbol{y}_s = (X_s'X_s)(X_s'X_s)^{-1}X_s'\boldsymbol{y}_s = (X_s'X_s)\hat{\beta}_{ols}$ (where $\hat{\beta}_{ols}$ is the OLS estimate of β), so its variance can easily be estimated as well.

Example and a Small Simulation

Stratified Random Sample
Consider the case of a stratified sample, where in each stratum h of size N_h $(h = 1, \ldots, H)$, a random sample of n_h units is drawn without replacement. Assume that *within stratum*, the sample selection process is conditionally independent of y_i, given x_i. In other words, in this example,

$$(5) \quad f(y_i|x_i, h_i, i \in s) = f(y_i|x_i, h_i).$$

The design variables in this case include the $H - 1$ stratum identifiers defined by $z_{h,i} = 1$ if unit i is in stratum h, $z_{h,i} = 0$ otherwise $(h = 1, \ldots, H - 1)$. Because of (5), we do not need to include other design variables as covariates. In this case, $A = (X_u'X_u)^{-1}X_u'(X_u, Z_u, D_xZ_u) = (I_2, KG, KH)$, where

$$(6a) \quad K = \begin{pmatrix} N & \sum_U x_i \\ \sum_U x_i & \sum_U x_i^2 \end{pmatrix}^{-1}$$

and where

$$(6b) \quad G = \begin{pmatrix} N_1 & \cdots & N_{H-1} \\ \sum_{U_1} x_i & \cdots & \sum_{U_{H-1}} x_i \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} \sum_{U_1} x_i & \cdots & \sum_{U_{H-1}} x_i \\ \sum_{U_1} x_i^2 & \cdots & \sum_{U_{H-1}} x_i^2 \end{pmatrix}.$$

Small Simulation Study
Following Pfeffermann (forthcoming), we simulate a population as follows.

1. The x variables were generated from a Gamma$(1, 2)$ distribution; $N = 1{,}000$ observations were generated. Each value was rounded to the nearest integer in the interval $[0, 5]$.

2. The units were partitioned to five strata $U_h, (h = 1, \ldots, 5)$ by means of independent draws for each unit i from Multinomial$(1, \boldsymbol{p})$ where $\boldsymbol{p} = (p_1, \ldots, p_5)'$, and $p_h = \Pr(i \in U_h) = \exp(b_h) \Big/ \sum_{k=1}^{5} \exp(b_k)$, and where $b_h = \frac{1}{3}(h-1), h = 1, 2, 3, 4, 5$. (Note that $b_1 = 0$ and so $h = 1$ is the reference level.) Let h_i be the stratum of unit i. Note that h_i is random.

3. The response variable y_i is independently drawn as

   $$y_i = 2 + x_i + (1 + 0.2x_i)\xi_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, 1),$$

   where $\xi_i = 0.2 \sum_{h=1}^{5} z_{h,i}/a_h(x_i) - 1$. Note that $E(\xi_i) = 0$ and $\text{Var}(\xi_i) = 0.2^2 \sum_{h=1}^{5} 1/a_h(x_i) - 1$ and therefore, $E(y_i|x_i) = 2+x_i$ and $\text{Var}(y_i|x_i) = (2+x_i)^2\text{Var}(\xi)+1 = (2+x_i)^2(0.2^2 \sum_{h=1}^{5} 1/a_h(x_i) - 1) + 1$. Also note that the random component of y_i, $(1+0.2x_i)\xi_i + \varepsilon_i$ is not normally distributed.

4. A measure of size was defined as $m_i = \max(\min(|z_i|^{1.5}, 9), 1)$, where $z_i \sim N(1 + x_i, 1)$.

5. Samples of sizes 60 each were drawn from each stratum without replacement and with systematic proportional to size. Thus the total sample size was $5 \times 60 = 300$.

6. One thousand finite populations were generated according to the scheme above, repeating steps 2, 3 and 4. Step 1 was performed only once. In other words, the x_i's were common to all the populations, while the stratum indicators $z_{h,i}$, and the response variable y_i were generated again for each of the 3,000 finite populations, and then a single sample was drawn from each new population.

Note that this set up defines an informative sampling design, i.e., the selection probabilities and the model outcome variable are related. This is due to the stratification and the model independent variable being related, and the unequal selection probabilities across the strata. Also note that the measure of size m_i is a function of x_i alone, and therefore $f(y_i|x_i, h_i, m_i) = f(y_i|x_i, h_i)$. Therefore, only the stratum indicator need to be included in the model in our proposed method
We considered the following estimation methods: (a) ordinary least squares (OLS), (b) probability weighted least squares (PWLS), (c) the proposed approach, with unknown x, z population values.
The model (4) fitted by the method Gelman (2007) is in general non-linear and does not produce estimates of the regression coefficients. Figure 1 below shows an example of one such fitted model (the piecewise line) and the 'true' model (the straight line). Therefore, its performance could not be directly compared to these methods. We did, however compare it with the other methods by the following measures:

$$\text{WRMSE} \overset{\text{def}}{=} \sqrt{(\sum_{i \in s} w_i)^{-1} \sum_{i \in s} w_i(\hat{y} - y_i)^2},$$
$$\text{WMAE} \overset{\text{def}}{=} (\sum_{i \in s} w_i)^{-1} \sum_{i \in s} |\hat{y} - y_i)|,$$

The WRMSE measure is an estimates of the finite population quantity $\sqrt{N^{-1} \sum_{i \in U} (\hat{y} - y_i)^2}$. Because PWLS is designed to minimize the WRMSE measure, we've included also the weighted mean absolute errors measure WMAE.

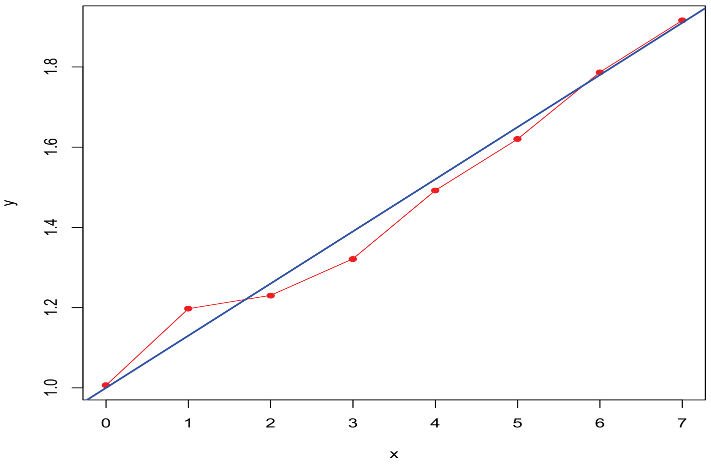

Figure 1: Example of Fitted Model by Gelman (2007) Method

Simulation Results

In Table 1 below the mean of the point estimates for both regression coefficients, the empirical standard errors, and the square roots of the variance estimates, both model-based and randomization-based, are shown. For estimating the variance of an estimator $\hat{\beta}_q = (X_s' Q_s X_s)^{-1} X_s' Q_s \boldsymbol{y}_s$ of the linear regression coefficient, where $Q_s = \text{diag}(q_1, \ldots, q_n)$ is a diagonal matrix of weights, two estimators can be used: (1) a model-based estimator given by $\hat{v}_m(\hat{\beta}_q) = (X_s' Q_s X_s)^{-1} X_s' Q_s^2 \text{diag}(e_i^2) X_s (X_s' Q_s X_s)^{-1}$, where the e_i are the residuals, X_s is the matrix whose rows are the sample units x variables (including 1 for the intercept), and (2) a randomization variance estimator given by $\hat{v}_r(\hat{\beta}_q) = (X_s' Q_s X_s)^{-1} \hat{v}_r(\hat{E}) (X_s' Q_s X_s)^{-1}$, where $\hat{v}_r(\hat{E})$ is the randomization (co)variance estimate of the estimated total, $\hat{E} = \sum_i q_i \boldsymbol{x}_i e_i$.

Table 1: Estimation of the regression coefficients

Method	β_0	β_1	$\text{SE}_{emp}(\beta_0)$	$\text{SE}_{emp}(\beta_1)$	$\bar{se}_m(\beta_0)$	$\bar{se}_m(\beta_1)$	$\bar{se}_r(\beta_0)$	$\bar{se}_r(\beta_1)$
Super population	2.000	1.000	0.000	0.000	-	-	-	
Census	2.000	1.000	0.026	0.012	-	-	-	
OLS	2.240	1.049	0.134	0.049	0.137	0.048	0.133	0.048
PWLS	2.000	1.000	0.167	0.059	0.166	0.055	0.164	0.054
Proposed	2.000	1.001	0.140	0.053	0.124	0.040	0.124	0.040

As expected, the OLS method is biased. Both the PWLS and the proposed method are unbiased. However, the proposed method is more efficient, as evident from the empirical standard errors. This is due to the measure of size m_i being a function of x_i alone, so that $f(y_i|x_i, h_i, m_i) = f(y_i|x_i, h_i)$, and therefore not required in the proposed method. In contrast, use of the weights in the PWLS adds to its variance. Note that the variance estimates for the proposed method are too low. This is probably due to ignoring the fact that the unobserved covariates and design variables were sampled from the observed data, and the added variance due to that was not accounted for.

As mentioned above, the Gelman(2007) method does not yield estimated regression coefficients. This Table 2 below compared the three methods and Gelman's method based on their WRMSE and WMAE.

Table 2: Prediction of the response variable

	WRMSE	WMAE
OLS	1.256	1.101
PWLS	1.206	0.959
Gelman	1.209	0.961
Proposed Method	1.208	0.961

Acknowledgement

This work is part of a project funded by an ESRC grant RES-062-23-2316 led by Danny Pfeffermann. His invaluable help is greatly appreciated.

REFERENCES

Gelman, A. (2007), “Struggles with survey weighting and regression modeling.” *Statistical Science*, Vol. **22**, 153–164.

Pfeffermann, D. (forthcoming), “Modelling of complex survey data: Why is it a problem? How can we approach it?”

Pfeffermann, D. and Sikov, A. (2011) “Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information.” *Journal of Official Statistics*, 181–209.

Pfeffermann, D. and Sverchkov, M. (1999), “Parametric and semi-parametric estimation of regression models fitted to survey data.” *Sankhya*, vol. **61**, 166–186.

Pfeffermann, D. and Sverchkov, M. (2009), “Inference under informative sampling.” In ‘Handbook of Statistics 29B; *Sample Surveys: Inference and Analysis*. Eds. D. Pfeffermann and C.R. Rao. North Holland. pp. 455–487.

Rubin, D. (1985), “The use of propensity scores in applied Bayesian inference.” Bayesian Statistics 2, eds. J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith, Elsevier Publishers, pp 463–472.

Skinner, C.J. (1994), “Sample Models and Weights.” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 133–142.