Journal of Data Science **5**(2007), 535-554

Distribution-Free Regression: Reinterpreting Design-Based Sampling

Gordon G. Bechtel
University of Florida and Florida Research Institute

Abstract: An individual in a finite population is represented by a random variable whose expectation is linearly composed of explanatory variables and a personal effect. This expectation locates her (his) random variable on a scale when s(he) responds to a questionnaire item or physical instrument. This formulation reinterprets design-based sampling, which represents an individual as a constant waiting to be observed. Retaining *constant expectations* , however, along with *fixed realizations* of random variables, preserves and strengthens design-based theory through the Horvitz-Thompson (1952) theorem. This interpretation reaffirms the usual design-based regression estimates, whose normality is seen to be free of any assumptions about the distribution of the outcome variable. It also formulates response error in a way that renders a superpopulation, postulated by model-based sampling, unnecessary. The value of distribution-free regression is illustrated with an analysis of American presidential approval.

Key words: Arbitrarily distributed response errors, latent versus manifest estimation, normally distributed estimates, sampling realizations, stochastic versus sampling variation, variance estimation, well-defined hidden variables.

1. A Paradigm Shift for Survey Sampling

Design-based sampling postulates an individual in a fixed observable state that may be subjectively measured as a discrete rating, e.g. 0 1 2 3 4, or physically measured, say, on a continuous blood pressure scale 0 ... 300mmHg. Thus, the value recorded in a survey interview or clinical trial is regarded as a fixed number in waiting. More realistically, however, an individual may be represented as a random variable that is realized in response to a questionnaire item or physical instrument.

The present paper favors this more plausible interpretation and extends Bechtel's (2005) treatment of survey proportions to survey regressions of any discrete or continuous dependent variable. The individual is posited here as a pair of fixed parameters; namely, a mean and variance that (partially) determine an idiosyncratic probability distribution. Each mean is composed of individual-specific

explanatory values along with an individual effect. Thus, a population of N individuals generates realizations of N non-identical distributions. Each individual realization is a momentary numerical value governed by an idiosyncratic mean and variance (and perhaps higher moments).

This approach brings the Horvitz-Thompson theorem to bear on a sample (without replacement) of n realizations from a momentary population of N realizations. It also enhances design-based regression, whose intercept and slopes are normally distributed (over samples) for *any* idiosyncratic distributions (over realizations) that prevail when humans respond to survey items and instruments.

Section 2 describes doubly bounded random variables, and section 3 lays out a survey regression that "explains" their expectations. Section 4 demonstrates the asymptotic normality of the estimated regression effects in the presence of any distributions that underlie the survey responses. Section 5 completes the present formulation with a treatment of variance estimation in this new context.

This paradigm is then applied in Section 6 to the important case of the population mean, which is simply a regression intercept in the absence of explanatory variables. Section 7 describes STATA commands for computing a distribution-free regression, and Section 8 illustrates this computation with American polling data. Section 9 sums up renewed design-based regression, noting its broad reach across opinion polling, economic surveys, and clinical trials

2. Individuals as Random Variables

2.1 Stochastic response error

For each individual $i = 1, \ldots, N$ in a population let Y_i be a random variable such that

$$Y_i = \eta_i + E_i$$

where E_i is a response error with

$$\begin{aligned} \mathbf{E}(E_i) &= 0, \\ \mathbf{Var}(E_i) &= \sigma_i^2, \quad \text{for } i = 1, 2, \ldots, N. \end{aligned}$$

The expectation and variance of the E_i are understood to be *over realizations* of non-identically distributed random errors $E_1, \ldots, E_N$.

In a survey the random variable Y_i may take the discrete values 0 1 2 3 4 5 6 for the responses *terrible, unhappy, mostly dissatisfied, mixed, mostly satisfied, pleased*, or *delighted* to a question about life quality (Andrews and Withey, 1976). The assumption of equal spacing between response labels, used in this paper, is venerable and ubiquitous (Galton, 1883; Thurstone, 1925; Likert, 1932; Coombs, 1964, pp. 211-212; Levy and Guttman, 1975; Clogg, 1979). Its robustness is

demonstrated in Section 8, where one (binary) response step produces regression slopes equivalent to those given by three equally spaced response steps.

In a clinical trial Y_i may take any value in the interval $0 \ldots 300$mmHg on a blood pressure scale. This continuous random variable, like the preceding discrete one, generates a measurement Y_i that departs from individual i's true value η_i. The expectation η_i denotes a personal location on the response scale, and the standard deviation σ_i (over realizations) denotes a personal uncertainty. For example, in a public opinion survey a low σ_i implies the crystallization of one's attitude. In a clinical trial a low σ_i denotes stability (or consistency) of one's blood pressure.

2.2 A linear characterization of η_i

Let $\boldsymbol{\eta}_p = [\eta_1, \ldots, \eta_N]^T$ be the vector of expectations in population $\mathbf{p}$ to be "explained" by an $N \times (k+1)$ population matrix $\mathbf{X}_p$. The vector $\boldsymbol{\eta}_p$ and matrix $\mathbf{X}_p$ *define* the finite population characteristic

$$\boldsymbol{\beta} = (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T \boldsymbol{\eta}_p \tag{2.1}$$

which is the target parameter here. The function $\boldsymbol{\beta}(\mathbf{X}_p, \boldsymbol{\eta}_p)$ in (2.1) then *defines*

$$\mathbf{A}_p = \boldsymbol{\eta}_p - \mathbf{X}_p \boldsymbol{\beta}, \tag{2.2}$$

which is a population vector $[\alpha_1, \ldots, \alpha_N]^T$ of residual idiosyncratic effects on $[\eta_1, \ldots, \eta_N]^T$ over and above $\mathbf{X}_p$. The η_i and α_i are well-defined hidden variables in present analysis.

In this setup, then, each individual $i = 1, \ldots, N$ in population $\mathbf{p}$ is represented as

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \alpha_i, \tag{2.3}$$

where η_i is i's expected response to a survey instrument, $X_{1i}, \ldots, X_{ki}$ are i's values of k variables that carry η_i through their effects $\beta_1, \ldots, \beta_k$, α_i is i's residual effect on η_i.

3. Sampling from One Realization

Assume a *single* realization $\{Y_1, \ldots, Y_N\}$ of the N random variables in Section 2.1. Observations $Y_1, \ldots, Y_n$ are now sampled (without replacement) from $\{Y_1, \ldots, Y_N\}$. This implies that response errors $E_1, \ldots, E_n$ are simultaneously sampled (without replacement) from the *one* realization $\{E_1, \ldots, E_N\}$. This setup reinterprets conventional design-based sampling which treats $\{Y_1, \ldots, Y_N\}$ and $Y_1, \ldots, Y_n$ as constants rather than realizations of random variables.

3.1 Weighting data in the presence of nonresponse

Missing data for *all* variables in $Y_i, X_{1i}, \ldots, X_{ki}$, called **unit nonresponse**, results in an absent survey protocol. For $i = 1, \ldots, N$ let π_i be the probability of i's inclusion in a selected sample and ϕ_i be the probability of i's survey participation given s(he) has been drawn. Regarding i's participation as the last (self-selection) stage of sampling, the probability that s(he) is in the subsample of n *observed realizations* is $\pi_i\phi_i$. In the sequel sample $\mathbf{s}$ denotes this subsample of n survey participants, each with case weight $w_i = 1/(\pi_i\phi_i)$. This case weight adjusts the sample design weight $1/\pi_i$ upward by the factor $1/\phi_i$ to compensate for any population under-representation in the sample $\mathbf{s}$ of observed realizations (Särndal and Lundström, 2005, pp. 43-44, 49-53).

The present reinterpretation of design-based sampling holds strictly under true case weights $w_i = 1/(\pi_i\phi_i)$. The probability ϕ_i (unlike π_i), however, is not known and is usually estimated for each unit in sample $\mathbf{s}$ by "weighting class" or "poststratification" adjustments (Lohr, 1999, pp. 264-272). If these n estimates approximate the true participation probabilities ϕ_i, the formulas below give nearly unbiased regression coefficients when the number of respondents n is large. Nevertheless, Lohr (1999, p. 272) cautions that "Weights may improve many of the estimates, but they rarely eliminate all nonresponse bias."

Missing data for *some* variables in $\{Y_i, X_{1i}, \ldots, X_{ki} : i \in \mathbf{s}\}$, called **item nonresponse**, gives an incomplete protocol for individual i in a survey regression. Various imputation procedures are available for filling in missing values in incomplete protocols (Lohr, 1999, pp. 272-278). However, because the theory here assumes that all responses have been realized in $\{Y_i, X_{1i}, \ldots, X_{ki} : i \in \mathbf{s}\}$, imputation adds bias to the regression formulas below . Lohr (1999, p. 277) notes, "If the nonresponse is missing at random given the covariates used in the imputation procedure, imputation substantially reduces the bias due to item nonresponse". Section 8.3 uses a regression imputation that avoids the loss of 29% of the cases due to missing item data in an American national survey (StataCorp., 2001, Volume 2, pp. 69-73; Särndal and Lundström, 2005, pp. 153-155, 158-161).

3.2 Estimating the target parameter $\boldsymbol{\beta}$

Each element of the $(k+1) \times (k+1)$ matrix $\mathbf{X}_p^T\mathbf{X}_p$ is a population sum of products, as is each element of the $(k+1) \times 1$ vector $\mathbf{X}_p^T\boldsymbol{\eta}_p$ (Lohr, 1999, p. 360). Therefore, due to Horvitz and Thompson (1952), unbiased estimates of these matrices are given by $\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\mathbf{X}_\mathbf{s}$ and $\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\boldsymbol{\eta}_\mathbf{s}$, where $\mathbf{X}_\mathbf{s}$ is the known $n \times (k+1)$ matrix of explanatory values in sample $\mathbf{s}$, $\boldsymbol{\eta}_\mathbf{s} = [\eta_1, \ldots, \eta_n]^T$ is the unknown respondent vector of individual expectations, $\mathbf{W}_\mathbf{s}$ =diag$(w_1, w_2, \ldots, w_n)$ is the known $n \times n$ diagonal matrix of case weights.

In a large sample $\mathbf{s}$ the *unobserved* Horvitz-Thompson (HT) estimator

$$\mathbf{b} = (\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\mathbf{X}_\mathbf{s})^{-1}\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\boldsymbol{\eta}_\mathbf{s} \tag{3.1}$$

is consistent and almost unbiased for $\boldsymbol{\beta}$, its unbiasedness being approximate because $\mathbf{b}$ is the product of estimators (Binder, 1983; Nathan, 1988, pp. 255-256; Thompson, 1997, pp. 106-107; Valliant, Dorfman, and Royall, 1999, pp. 40-41; Lohr, 1999, pp. 354-361). Similarily, for the realized (but unobserved) response error $\mathbf{E}_\mathbf{s} = [E_1, \ldots, E_n]^T$ in $\mathbf{s}$ the *unobserved* HT estimator

$$\mathbf{v} = (\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\mathbf{X}_\mathbf{s})^{-1}\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\mathbf{E}_\mathbf{s} \tag{3.2}$$

is consistent and almost unbiased for $\boldsymbol{\upsilon} = (\mathbf{X}_\mathbf{p}^T\mathbf{X}_\mathbf{p})^{-1}\mathbf{X}_\mathbf{p}^T\mathbf{E}_\mathbf{p}$. The vector $\mathbf{E}_\mathbf{p} = [E_1, \ldots, E_N]^T$ consists of the realized response errors in $\mathbf{p}$ which are transformed to $\boldsymbol{\upsilon}$. Also, as seen in (3.5) below, the error transform $\mathbf{v}$ in (3.2) delivers respondent error $\mathbf{E}_\mathbf{s}$ to the manifest regression effects.

Finally, for the realized and observed measurement $\mathbf{Y}_\mathbf{s} = [Y_1, \ldots, Y_n]^T$ in $\mathbf{s}$ the *observed* HT estimator

$$\mathbf{B} = (\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\mathbf{X}_\mathbf{s})^{-1}\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\mathbf{Y}_\mathbf{s} \tag{3.3}$$

is consistent and almost unbiased for

$$\boldsymbol{\theta} = (\mathbf{X}_\mathbf{p}^T\mathbf{X}_\mathbf{p})^{-1}\mathbf{X}_\mathbf{p}^T\mathbf{Y}_\mathbf{p} \tag{3.4}$$

where $\mathbf{Y}_\mathbf{p} = [Y_1, \ldots, Y_N]^T$ consists of the realized measurements in $\mathbf{p}$. Formula (3.3) is the estimator of the conventional target (3.4) in design-based regression (Frankel, 1971, pp.7-25; Lohr, 1999, pp. 359-361; StataCorp., 2001, Volume 4, pp. 29-30; Chaudhuri and Stenger, 2005, pp. 264-265). However, in the present reinterpretation $\mathbf{B}$ also, and more profoundly, estimates the new target $\boldsymbol{\beta}$ in (2.1). Thus, given $\mathbf{Y}_\mathbf{s}$ as a subvector of the fixed vector $\mathbf{Y}_\mathbf{p}$ of realizations,

$$\mathbf{B} = \mathbf{C}_\mathbf{s}\mathbf{Y}_\mathbf{s} = \mathbf{C}_\mathbf{s}(\boldsymbol{\eta}_\mathbf{s} + \mathbf{E}_\mathbf{s}) = \mathbf{C}_\mathbf{s}\boldsymbol{\eta}_\mathbf{s} + \mathbf{C}_\mathbf{s}\mathbf{E}_\mathbf{s} = \mathbf{b} + \mathbf{v}, \tag{3.5}$$

where

$$\mathbf{C}_\mathbf{s} = (\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}\mathbf{X}_\mathbf{s})^{-1}\mathbf{X}_\mathbf{s}^T\mathbf{W}_\mathbf{s}$$

Taking expectations over samples $\mathbf{s}$ of size n then gives

$$\mathbf{E}(\mathbf{B}) = \mathbf{E}(\mathbf{b}) + \mathbf{E}(\mathbf{v}) \approx \mathbf{E}(\mathbf{b}) \approx \boldsymbol{\beta} \tag{3.6}$$

because

$$\mathbf{E}(\mathbf{v}) \approx \boldsymbol{\upsilon} = \boldsymbol{\theta} - \boldsymbol{\beta} \approx \mathbf{0}. \tag{3.7}$$

Therefore, $\mathbf{B}$ is almost unbiased for $\boldsymbol{\beta}$ in large-sample surveys.

Equations (2.1) through (3.7) have been developed from a single realization of N *arbitrary* random variables without reference to any superpopulation with an assumed probability density and covariance structure (cf. Skinner, Holt, and Smith, 1989; Thompson, 1997; Valliant, Dorfman, and Royall, 1999; Lohr, 1999; Binder and Roberts, 2003).

4. Normality of the Regression Effects

The expectation of (3.4) *over realizations* of the stochastic $\mathbf{Y_p}$ is

$$\begin{aligned} \mathbf{E}(\boldsymbol{\theta}) &= (\mathbf{X_p}^T\mathbf{X_p})^{-1}\mathbf{X_p}^T\mathbf{E}(\mathbf{Y_p}) \\ &= (\mathbf{X_p}^T\mathbf{X_p})^{-1}\mathbf{X_p}^T\boldsymbol{\eta_p} = \boldsymbol{\beta}. \end{aligned} \tag{4.1}$$

Because $E(\theta_j) = \beta_j$ and $\mathrm{Var}(\theta_j) \to 0$ as $N \to \infty$ for $j = 0, \ldots$,k, the difference $\theta_j - \beta_j = v_j$ is infinitesimal for a *given realization* of $\mathbf{Y_p}$. Fixing this momentary realization $\{Y_1, \ldots, Y_N\}$, the resulting reals $\theta_0, \ldots, \theta_k$ become the classic target parameters of design-based regression. Therefore, a strict design-based argument using the θ_j can be given for the normality over samples of each element B_j in $\mathbf{B}$. This provides a statistic for testing hypotheses about the target parameter β_j against the observed coefficient B_j.

First, given the realization $\{Y_1, \ldots, Y_N\}$, the coefficient θ_j $(j = 0, \ldots, k)$ can be written as a smooth function of population totals of cross products in $\{Y_i, 1, X_{1i}, \ldots, X_{ki} : i \in \mathbf{p}\}$. Then, from the subset $\{Y_i, 1, X_{1i}, \ldots, X_{ki} : i \in \mathbf{s}\}$ the estimate B_j can be written as the *same* function of HT estimators of these population totals. The HT estimators are corresponding sample totals of cross products with each term case weighted by w_i . For example, $\sum_{i\in\mathbf{s}} w_i X_{1i} Y_i$ is an HT estimator of $\sum_{i\in\mathbf{p}} X_{1i} Y_i$ (cf. Lohr, 1999, pp. 352-360; Thompson, 1997, pp. 106-108).

Next, a Taylor series "linearization" of the error

$$\epsilon(B_j; \theta_j) = B_j - \theta_j \approx \epsilon_j,$$

along with asymptotic multivariate normality of the HT estimators, implies that $(B_j - \theta_j)/\sqrt{\mathrm{Var}(\epsilon_j)}$ is asymptotically $N(0, 1)$ (Lehmann, 1999, pp. 253-269, 309-315; Lohr, 1999, pp. 290-293, 310, 352-360; Sen, 1988, pp. 313-328; Thompson, 1997, pp. 58-64, 106-111). The estimate $Var(\epsilon_j)$ of $\mathrm{Var}(\epsilon_j)$ is given in Section 5 and computed by software described in Section 7. Finally, due to the infinitesimal difference between θ_j and β_j, the statistic

$$t = (B_j - \beta_{j0})/\sqrt{Var(\epsilon_j)} \tag{4.2}$$

may be used to test an hypothesis $H : \beta_j = \beta_{j0}$ about our target coefficient β_j. In applying (4.2) it is reassuring to recall that the asymptotic normality of B_j

(over samples) does not depend on the distributions of $Y_1, \ldots, Y_N$ (over realizations). Large-sample normality of the estimated intercept and slopes prevails in the presence of *any* idiosyncratic distributions of survey responses.

5. Variance Estimation

Using "linearization" in design-based sampling, an estimate of the covariance matrix of $\mathbf{b}$ in (3.1) is given by

$$Var(\mathbf{b}) = (\mathbf{X_s}^T\mathbf{W_s}\mathbf{X_s})^{-1}Var[\sum_{i\in\mathbf{s}} w_i X_i^T(\eta_i - X_i^T\mathbf{b})](\mathbf{X_s}^T\mathbf{W_s}\mathbf{X_s})^{-1}, \qquad (5.1)$$

where w_i is the case weight of respondent i, $X_i^T = [1, X_{1i}, \ldots, X_{ki}]$ is row i of $\mathbf{X_s}$, and $\eta_i - X_i^T\mathbf{b}$ is a latent estimate of i's effect $\alpha_i = \eta_i - X_i^T\boldsymbol{\beta}$. The unobserved matrix $Var(\mathbf{b})$ in (5.1) is based on a strict application of design-based regression (cf. Lohr, 1999, pp. 359-361; StataCorp., 2001, Volume 4, pp. 29-30). Thus a vector $\boldsymbol{\eta_s} = [\eta_1, \ldots, \eta_n]^T$ is sampled from a population vector $\boldsymbol{\eta_p} = [\eta_1, \ldots, \eta_N]^T$ of *constants*, and $\mathbf{b}$ is computed from $\boldsymbol{\eta_s}$ using (3.1). It is not possible to observe the core residual $\eta_i - X_i^T\mathbf{b}$ in (5.1) because $\boldsymbol{\eta_s}$ is not observed. Hence $Var(\mathbf{b})$ is a latent estimate.

Next, noting that $\mathbf{E_s} = \mathbf{Y_s} - \boldsymbol{\eta_s}$ is "predicted" by

$$\mathbf{X_s}\mathbf{B} - \mathbf{X_s}\mathbf{b} = \mathbf{X_s}(\mathbf{B} - \mathbf{b}) = \mathbf{X_s}\mathbf{v},$$

the covariance matrix of $\mathbf{v}$ in (3.2) can be estimated like $Var(\mathbf{b})$. Replacing the core residual in (5.1) by the "residual" $E_i - X_i^T\mathbf{v}$ gives

$$Var(\mathbf{v}) = (\mathbf{X_s}^T\mathbf{W_s}\mathbf{X_s})^{-1}Var[\sum_{i\in\mathbf{s}} w_i X_i^T(E_i - X_i^T\mathbf{v})](\mathbf{X_s}^T\mathbf{W_s}\mathbf{X_s})^{-1}. \qquad (5.2)$$

Because the E_i are not observed, $Var(\mathbf{v})$ is also a latent estimate.

Simulated response errors E_i demonstrate that the sum of the latent estimates in (5.1) and (5.2) closely approximates

$$\begin{aligned} Var(\mathbf{B}) &= (\mathbf{X_s}^T\mathbf{W_s}\mathbf{X_s})^{-1}Var[\sum_{i\in\mathbf{s}} w_i X_i^T(Y_i - X_i^T\mathbf{B})](\mathbf{X_s}^T\mathbf{W_s}\mathbf{X_s})^{-1} \\ &\approx Var(\mathbf{b}) + Var(\mathbf{v}). \end{aligned} \qquad (5.3)$$

The variance estimator in (5.3) is identical to that in conventional design-based regression (Lohr, 1999, pp. 359-361; StataCorp., 2001, Volume 4, pp. 29-30). Its reinterpretation here is seen by writing i's manifest residual as

$$\begin{aligned} Y_i - X_i^T\mathbf{B} &= \eta_i + E_i - X_i^T(\boldsymbol{\beta} + \mathbf{u} + \mathbf{v}) \\ &= \alpha_i + E_i - X_i^T(\mathbf{u} + \mathbf{v}), \end{aligned} \qquad (5.4)$$

where $\mathbf{u} = \mathbf{b} - \boldsymbol{\beta}$ and $\mathbf{v} = \mathbf{B} - \mathbf{b}$. Equation (5.4) shows that i's observed residual equals her (his) actual residual in addition to a component containing two estimation errors; namely, the departure of the latent estimate $\mathbf{b}$ from $\boldsymbol{\beta}$ and the deviation of the manifest estimate $\mathbf{B}$ from $\mathbf{b}$.

Finally, rewriting (5.4) as

$$\begin{aligned} Y_i - X_i^T\mathbf{B} &= \eta_i + E_i - X_i^T(\boldsymbol{\beta} + \mathbf{u} + \mathbf{v}) \\ &= \eta_i + E_i - X_i^T(\mathbf{b} + \mathbf{v}) \\ &= (\eta_i - X_i^T\mathbf{b}) + (E_i - X_i^T\mathbf{v}) \end{aligned} \tag{5.5}$$

shows that this manifest residual also equals i's residual in (5.1) plus her (his) residual in (5.2). Thus $Var(\mathbf{b})$ expands to $Var(\mathbf{B})$ due to response error E_i and estimation error $\mathbf{v} = \mathbf{B} - \mathbf{b}$. In particular, the $k+1$ diagonals of $Var(\mathbf{B})$, which are the manifestly estimated variances of the observed $B_0, B_1, \ldots, B_k$, are larger than the $k+1$ diagonals of $Var(\mathbf{b})$. These latter diagonals are the latently estimated variances of the unobserved $b_0, b_1, \ldots, b_k$, which are generated by sampling $\eta_1, \ldots, \eta_n$ from the population $\{\eta_1, \ldots, \eta_N\}$.

6. The Important Case of the Mean

If the explanatory variables $X_{1i}, \ldots, X_{ki}$ are deleted from (2.3), then $\mathbf{X_p}$ becomes the unit vector containing N ones. In this special case the target parameter (2.1) is the *population mean expectation*

$$\beta_0 = N^{-1} \sum_{i \in \mathbf{p}} \eta_i$$

Correspondingly, substituting the unit vector of n ones for $\mathbf{X_s}$ in (3.1) gives the latent estimate

$$b_0 = \sum_{i \in \mathbf{s}} w_i \eta_i / \sum_{i \in \mathbf{s}} w_i,$$

which is *exactly* unbiased for β_0 because

$$\begin{aligned} \mathbf{E}\{\sum_{i \in \mathbf{s}} w_i \eta_i\} &= \sum_{i \in \mathbf{p}} \eta_i \qquad \text{and} \\ \sum_{i \in \mathbf{s}} w_i &= N. \end{aligned}$$

Next, substituting this unit vector for $\mathbf{X_s}$ in (3.2) and (3.3) gives

$$\begin{aligned} v_0 &= N^{-1} \sum_{i \in \mathbf{s}} w_i E_i \qquad \text{and} \\ B_0 &= \sum_{i \in \mathbf{s}} w_i Y_i / \sum_{i \in \mathbf{s}} w_i. \end{aligned}$$

The latter formula for B_0 is well-known in design-based sampling as the estimate of

$$\theta_0 = N^{-1} \sum_{i \in \mathbf{p}} Y_i$$

(Lohr, 1999, p.198; StataCorp., 2001, Volume 4, p.70). This target θ_0 is found here by substituting the unit vector of N ones for $\mathbf{X_p}$ in (3.4). In conventional design-based theory θ_0 is the mean of *fixed constants* $Y_1, \ldots, Y_N$. Here it is the mean of N *realizations* of random variables.

With the inclusion of response error in the present reinterpretation, B_0 is also seen to estimate β_0, which is the population mean of expectations $\eta_1, \ldots, \eta_N$. Thus using (3.5),

$$B_0 = b_0 + v_0,$$

and taking expectations over samples $\mathbf{s}$ of size n and using (3.6) and (3.7),

$$E(B_0) = E(b_0) + E(v_0) \approx E(b_0) = \beta_0$$

because $E(v_0) \approx 0$. Therefore, in large samples B_0 is almost unbiased for β_0 .

Finally, substituting the unit vector for $\mathbf{X_s}$ in (5.3) gives an estimate of the variance of B_0:

$$Var(B_0) = Var[\sum_{i \in \mathbf{s}} w_i(Y_i - B_0)]/[\sum_{i \in \mathbf{s}} w_i]^2.$$

Writing i's weighted residual $w_i(Y_i - B_0)$ as U_i, the estimated variance in the numerator of $Var(B_0)$ is computed as

$$Var[\sum_{i \in \mathbf{s}} U_i] = n \sum_{i \in \mathbf{s}} (U_i - \bar{U})^2/(n-1),$$

where $\bar{U} = \sum_{i \in \mathbf{s}} U_i/n$ (StataCorp., 2001, Volume 4, pp. 29-30, 70).

The formulas in this section show that the mean of a survey variable is the intercept of a distribution-free regression whose slopes are set to zero. In this special case too the intercept β_0, its latent estimate b_0, and its manifest estimate B_0 are defined without reference to a superpopulation.

7. Software for Distribution-Free Regression

The regression coefficients in (3.3), along with their standard errors from (5.3), are easily computed with two STATA commands:

$$\text{svyset pweight weight} \tag{7.1}$$

$$\text{svyreg } Y \ X_1 \ \ldots X_k \tag{7.2}$$

(StataCorp., 2001, Volume 4, pp.18-31). In (7.1) and (7.2) *weight* is a user-supplied variable containing case weights, Y is the survey response variable, and $X_1, \ldots, X_k$ are the predictors on which the regression is conditioned. This STATA setup returns the regression effects $B_0, B_1, \ldots, B_k$ and their standard errors. As shown by (5.4) and (5.5), the $k+1$ standard errors delivered by (7.1) and (7.2) reflect the effects of the response errors E_i on the variances of $B_0, B_1, \ldots, B_k$.

8. Opinion polling: American Presidential Approval

8.1 The quaternary regression model

This section uses survey data that sharply departs from the (usually assumed) continuity, normality, and homoscedasticity of the Y_i . The breakdown for our coded survey measure is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \alpha_i + E_i \\ &= X_i^T \boldsymbol{\beta} + \alpha_i + E_i \\ &= \eta_i + E_i \quad \text{for} \;\; i = 1, \ldots, n, \end{aligned} \tag{8.1}$$

where Y_i is i's *observed realization* on the integers 0 1 2 3, $X_{1i}, \ldots, X_{ki}$ are i's values on k explanatory variables, α_i is i's residual effect on her (his) η_i, $E_i = Y_i - \eta_i$ is i's *unobserved realized* error, and $Y_i - X_i^T \boldsymbol{\beta} = \alpha_i + E_i$ is i's regression residual. Equation (8.1) is estimated with $B_0, B_1, \ldots, B_k$ calculated from (3.3). This manifestly estimates i's regression residual as $Y_i - X_i^T \mathbf{B}$, whose latent components are given in (5.4) and (5.5).

8.2 The survey items and sample

The response values 0 1 2 3 taken by Y_i in (8.1) code four response options to the following item:

Overall, how would you rate President Bush's performance on the job?

Poor	Fair	Good	Excellent
(0)	(1)	(2)	(3)

This item is administered monthly by *Zogby International*, who monitors the perceived performance of the American President. Presidential approval is a closely watched variable that is also tracked by the Gallup Organization, CBS News/*New York Times*, ABC/*Washington Pos*t, NBC News/*Wall Street Journal*, and the American National Election Studies (Clarke, Stewart, and Rodgers, 2005).

This non-normal, discrete, and heteroscedastic variable was regressed on nine predictors also measured in the Zogby poll. Responses to these nine explanatory items in Table 1 are also coded 0 1 2 3. Therefore, the nine regression slopes in Tables 2 and 3 are comparable in magnitude. The *overall opinion* item in Table 1, along with the eight *specific performance* items, serve as mutual controls in predicting presidential job approval.

The 1009 respondents to these items were selected by probability sampling and contacted by computer-assisted telephone interviewing (CATI) between February 25 and 27, 2005. This was one month into the second presidential term of George W. Bush, who was reelected in the autumn of 2004. Case weights for the 1009 respondents were obtained from a demographic profile geared to the American population. These weights reflect region, political party, age, race, religion, and gender in order to more accurately represent this population.

Table 1: Predictors of presidential performance

Item	Response			
Overall opinion of George W. Bush	Very unfavorable	Somewhat unfavorable	Somewhat favorable	Very favorable
Jobs and the economy	Poor	Fair	Good	Excellent
The Iraq war	Poor	Fair	Good	Excellent
The environment	Poor	Fair	Good	Excellent
Foreign policy	Poor	Fair	Good	Excellent
Social security and Medicare	Poor	Fair	Good	Excellent
Education	Poor	Fair	Good	Excellent
Taxes	Poor	Fair	Good	Excellent
The war on terrorism	Poor	Fair	Good	Excellent

Source: This table is adapted from an SPSS data file provided by Zogby International.

8.3 Analysis and results

Missing rates for the predictors in Table 1 are 4% or less, except for *the environment, foreign policy*, and *taxes* which have 6%, 12%, and 13% missing responses. Because six of these rates are very low, and in order to preserve sample size, a regression imputation was carried out for each of the nine predictors against the other eight (cf. StataCorp., 2001, Volume 2, pp. 69-73).

Table 2: Quaternary Regression of presidential job performance ($R^2 = .72$)

Explanatory variable	Slope	t statistic	Probability
Overall opinion of George W. Bush	.282	7.44	.000
Performance on specific issues			
Jobs and the economy	.187	5.61	.000
The Iraq war	.184	4.80	.000
The environment	.165	4.99	.000
Foreign policy	−.149	−3.98	.000
Social security and Medicare	.128	3.19	.001
Education	.107	2.80	.005
Taxes	−.038	−1.05	.293
The war on terrorism	.036	.95	.343

Source: The values in this table were obtained from the STATA commands svyset and svyreg described in the text. This linear survey regression was carried out on a STATA spreadsheet translated from an SPSS data file supplied by Zogby International. The translation was done with STAT/TRANSFER software obtained from *Circle Systems, Inc.*

Using the STATA commands in (7.1) and (7.2), 979 non-missing ratings of George W. Bush's *job performance* Y were regressed on the imputed predictors $X_1, \ldots, X_9$. The dependent variable Y was not imputed due to its low missing rate of 3%. The nine estimated slopes $B_1, \ldots, B_9$ are exhibited in Table 2, where the predictors are ranked in the order of their effects on perceived job performance. As already noted, these slopes are comparable in magnitude due to the 0 1 2 3 coding of $X_1, \ldots, X_9$.

The R^2 of .72 indicates that almost three-quarters of the variance in presidential job approval is explained by these nine predictors in the Zogby poll. *Overall favorability* toward George W. Bush is the strongest predictor. Controlling for this general opinion, the quaternary regression also shows that *jobs and the economy* and *the Iraq war* are the most specifically predictive of overall job performance. (These two issues remain paramount for the American public at the present writing.) *The environment, foreign policy, social security and Medicare*, and *education* show an evenly descending gradient in the strength of their regression effects. *Foreign policy*, surprisingly, is negative in sign suggesting that the American public looks unfavorably on presidential efforts in this direction. Finally, Table 2 shows that in February 2005 *taxes* and *the war on terrorism* were unimportant issues. This despite the administration's emphasis on the importance of lowering taxes and its asserted link between its wars on terrorism and

Iraq.

Table 3: Binary Regression of presidential job performance ($R^2 = .62$)

Explanatory variable	Slope	t statistic	Probability
Overall opinion of George W. Bush	.123	6.30	.000
Performance on specific issues			
Jobs and the economy	.079	4.38	.000
The Iraq war	.078	3.80	.000
The environment	.059	3.31	.001
Foreign policy	−.060	−2.65	.008
Social security and Medicare	.046	2.22	.026
Education	.058	2.88	.004
Taxes	−.003	−.14	.891
The war on terrorism	.014	.65	.519

Source: The values in this table were obtained from the STATA commands *svyset* and *svyreg* described in the text. This linear survey regression was carried out on a STATA spreadsheet translated from an SPSS data file supplied by Zogby International. The translation was done with STAT/TRANSFER software obtained from *Circle Systems, Inc.*

8.4 A binary regression

Departures from continuity, normality and homoscedasticity for Y_i in (8.1) are now pressed to the most extreme case in which Y_i is dichotomous on the integers 0 1. This alternative dependent variable was generated by recoding the Zogby data as follows:

Overall, how would you rate President Bush's performance on the job?

Poor Fair Good Excellent

Negative Positive

(0) (1)

The resulting 979 binary measures were also regressed on the nine imputed predictors in Table 1 using the STATA commands in (7.1) and (7.2). The nine regression slopes, exhibited in Table 3, are plotted on their quaternary counterparts in Figure 1. The near perfect linearity (through the origin) of this plot demonstrates that distribution-free regression delivers valid slopes, even in its most

extreme case of *one step* from "negative" to "positive". Conversely, as noted in Section 2.1, the equivalence of these quaternary and binary slopes demonstrates the robustness of assuming three equal steps from "poor" to "fair" to "good" to "excellent". Tables 2 and 3 show that quaternary slopes enjoy larger t statistics and a greater R^2 than binary slopes. Evidently quaternary regressions are to be preferred, especially since they also offer an easy choice task to the respondent.

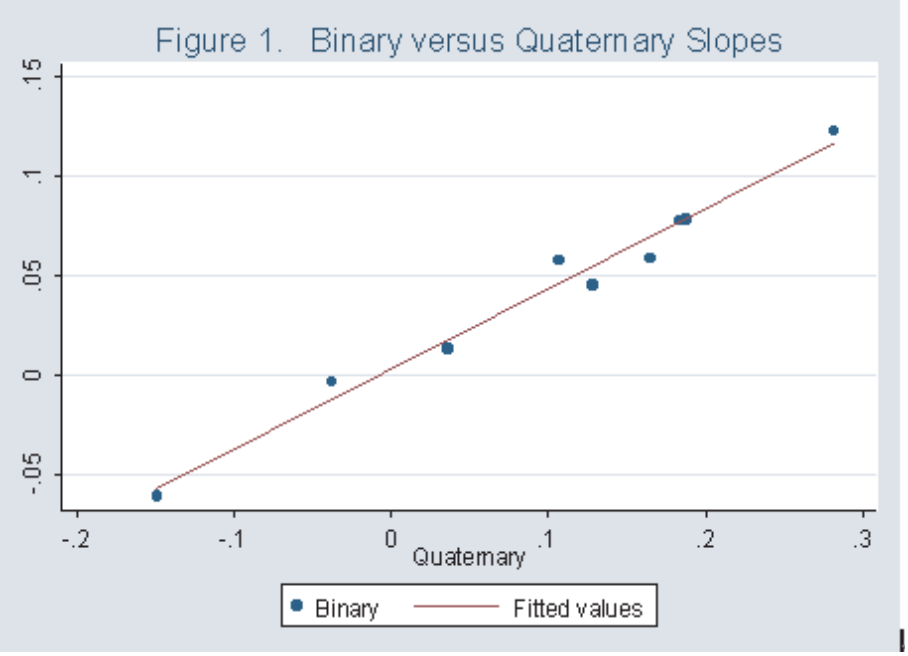

Figure 1: Binary versus quaternary slopes

9. The Reach of Renewed Design-based Regression

The present work replaces a population of constants with a population of random variables. Both of these populations produce an observed sample of numbers, but their generating processes are very different. In conventional design-based sampling, fixed individual states are believed to be selected and observed directly. In the present reinterpretation, *stochastic response error* generates N individual random variables that are realized in a population. Subsequently, n of these realizations are observed in a sample from this population. The status-quo theory (unrealistically) regards these population and sample realizations as fixed and immutable constants (cf. Lehmann, 1999, pp.115-116).

9.1 Binary responses

Bechtel (2005) introduced the distinction between these two types of populations in binary applications. In this case a population of Bernoulli variates, governed by personal probabilities, replaces a population of fixed 0's and 1's. The sample proportion produced by survey solicitations estimates the population mean of these probabilities, which are individual *expectations* of responding 0 or 1. In contrast, conventional design-based sampling interprets this same sample proportion as the mean of a population of 0s and 1s that are *fixed and noticeable states*.

The application in Section 8.4 extends Bechtel (2005) by analysing personal probabilities η_i with equation (8.1). Here E_i is a binary response error causing η_i to manifest as $Y_i = 0$ or $Y_i = 1$. Even with this most extreme departure of survey data from continuity, normality, and homoscedasticity, normality (over samples) of *regression effects* on personal probabilities is justified in Section 4.

9.2 Equal-step response scales

The regression of binary variates generalizes immediately to discrete random variables that code ordered responses in public opinion polls. This type of survey item solicits individual choice behavior (cf. Luce, 1959) over a set of options such as *poor, fair, good*, and *excellent* in Section 8.2. There E_i is a quaternary response error causing i's expectation η_i to manifest as $Y_i = 0$ 1 2 or 3. In the case of three response options, such as *disagree, neutral*, and *agree*, i's expectation η_i is continuous but her (his) potential realizations Y_i are limited to the integers 0 1 or 2.

Discrete dependent variables may also arise from multiple-item scales in survey questionnaires. For example, the ternary coding 0 1 2 may be used for each of three items that measure a subjective attribute. Summing these three item scores generates a discrete random variable Y_i whose expected value for individual i is

$$\eta_i = \eta_{i1} + \eta_{i2} + \eta_{i3}.$$

The true value η_i is continuous, whereas Y_i is restricted to the integers 0 1 2 3 4 5 or 6. The assumption in Section 2.1 that $Y_i = \eta_i + E_i$, i.e. that a fixed individual i's observed scale score equals her (his) true score plus a random error, has been used in psychological test theory by Lord and Novick (1968, pp. 27-38). The variance of Y_i (over realizations) on the three-item scale in the present illustration is

$$\sigma_i^2 = \sigma_{i1}^2 + \sigma_{i2}^2 + \sigma_{i3}^2 + 2\gamma_{i12} + 2\gamma_{i13} + 2\gamma_{i23},$$

where, for example, γ_{i12} is the covariance of i's responses to items 1 and 2. These

inter-item covariances relax the implausible assumption of "local independence" in item response theory, which requires that

$$\gamma_{i12} = \gamma_{i13} = \gamma_{i23} = 0,$$

i.e. that individual i's responses to successive questionnaire items be independent. (Embretson and Reise, 2000).

Single and multiple-item scales have been a mainstay of psychological measurement since Sir Francis Galton (1883) first introduced the rating of subjective attributes. Coombs (1964, pp. 211-212) gave various reasons for the ubiquity of rating scales, whose common use dates back to the early 1900s (Thurstone, 1925; Likert, 1932). In the 1970s survey ratings underlaid the measurement of life quality. This effort was stimulated by Levy and Guttman (1975), Andrews and Withey (1976), and Clogg's (1979) latent class analysis of the 1975 General Social Survey. In that decade the rating scale was also the vessel for consumer satisfaction (Bechtel, 1977). Subsequently the quality-control revolution, stimulated by the earlier work of W. Edwards Deming (Mann, 1994), led to worldwide preoccupation with satisfaction. In the public and private sectors this concern surfaced as "outcome evaluation", where rated satisfaction is solicited in national surveys and clinical trials.

9.3 Continuous scales

In contrast to discrete scales for measuring subjective variables, clinical and economic measures tap physical properties such as blood pressure and wealth. Here too status quo sampling theory unrealistically postulates fixed blood pressures in the population, rather than realizations of individual random variables. The alternative here samples these realizations which are continuous in mmHg units. Each reading Y_i departs from i's true pressure η_i due to response error E_i. Instead of being equally spaced these Y_i, like their η_i, are continuous on the scale $0 \ldots 300$mmHg.

9.4 Explaining individual idiosyncrasies

In both its discrete and continuous applications reinterpreted design-based theory represents respondent i as an idiosyncratic probability distribution. Her (his) random variable Y_i differs from its true value η_i due to a stochastic response error E_i defined in Section 2.1. A continuous Y_i, along with its mean η_i, can take any value on the response scale. Discrete Y_i, however, are restricted to equally spaced response values. In Section 8.2 the values 0 1 2 3 code the well known Zogby scale of *poor, fair, good*, or *excellent* presidential performance.

When $Y_1, \ldots, Y_n$ are sampled from a population realization $\{Y_1, \ldots, Y_N\}$, the regression estimate $\mathbf{B}$ in (3.3) is asymptotically normal over samples $\mathbf{s}$ and almost unbiased for $\boldsymbol{\beta}$ in (2.1). The target $\boldsymbol{\beta}$ partially accounts for response expectations $\eta_1, \ldots, \eta_N$ in a finite population of individuals. The variances of coefficients $B_0, B_1, \ldots, B_k$ in $\mathbf{B}$ are estimated by the diagonals of the matrix in (5.3). These same diagonals are used in conventional design-based theory, where $Y_1, \ldots, Y_n$ are (implausibly) regarded as drawn from a population $\{Y_1, \ldots, Y_N\}$ of human *constants*. Alternatively, the renewed theory here interprets each Y_i as a realization of a random variable (partially) governed by i's personal parameters η_i and σ_i^2 defined in Section 2.1. This interpretation better justifies formulas (3.3) and (5.3), long used to estimate survey regression coefficients and their standard errors. It also strengthens the foundation of design-based theory by realistically representing human populations as finite sets of *unique* individuals who are subject to idiosyncratic response errors. The errors considered here occur in the absence of a hypothetical superpopulation with particular distribution and covariance structures. These arbitrarily distributed errors lend credibility to the widely-used formulas of design-based regression theory.

Acknowledgement

This work has been supported by the University of Florida's Warrington College of Business. The author thanks the JDS editor and referee whose comments have improved the paper. Thanks also go to John Zogby and Christopher Conroy of Zogby International who kindly provided the data used to illustrate distribution-free regression. Of course none of the ideas or analyses here are attributable to the University of Florida or Zogby International.

References

Andrews, F. M. and Withey, S. B. (1976). *Social Indicators of Well-Being*. Plenum Press.

Bechtel, G. G. (1977). A model for monitoring consumer satisfaction. In *Conceptualization and Measurement of Consumer Satisfaction and Dissatisfaction* (Edited by H. K. Hunt), 187-214. Marketing Science Institute.

Bechtel, G. G. (2005). Sampling random variables: A paradigm shift for opinion polling. *Journal of Data Science* **3**, 439-448.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex samples. *International Statistical Review* **51**, 279-292.

Binder, D. A. and Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. In *Analysis of Survey Data* (Edited by R.L. Chambers and C.J. Skinner), 29-48. Wiley.

Chaudhuri, A. and Stenger, H. (2005). *Survey Sampling: Theory and Methods*, 2nd edition. Chapman and Hall.

Clarke, H. D., Stewart, M. C. and Rodgers, C. (2005). Presidential approval. In *Polling America: An Encyclopedia of Public Opinion* (Edited by S.J. Best and B. Radcliff), 571-579, Greenwood Press.

Clogg, C. C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research* **8**, 287-301.

Coombs, C. H. (1964). *A Theory of Data*. Wiley.

Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Erlbaum.

Frankel, M. R. (1971). *Inference from Survey Samples: An Empirical Investigation*. A revised version of a University of Michigan doctoral dissertation.

Galton, F. (1883). *Inquiries into Human Faculty and its Development*. Macmillan.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer.

Levy, S. and Guttman, L. (1975). On the multivariate structure of well-being. *Social Indicators Research* **2**, 361-388.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology* No. 140, 1-55.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Lord, F. M. and Novick, M, R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Luce, R. D. (1959). *Individual Choice Behavior*. Wiley.

Mann, N. R. (1994). W. Edwards Deming 1900-1993. *Journal of the American Statistical Association* **89**, 365-366.

Nathan, G. (1988). Inference based on data from complex sample designs. In *Handbook of Statistics, Volume 6 (sampling)* (Edited by P. R. Krishnaiah and C. R. Rao), 247-266. North-Holland.

Särndal, C. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley.

Sen, P. K. (1988). Asymptotics in finite population sampling. In *Handbook of Statistics, Volume 6 (Sampling)* (Edited by P. R. Krishnaiah and C. R. Rao), 291-331, North Holland.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. Wiley.

StataCorp. (2001). Stata Statistical Software: Release 7.0. Stata Corporation.

Thompson, M. E. (1997). *Theory of Sample Surveys*. Chapman & Hall.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology* **16**, 433-451.

Valliant, R., Dorfman, A. H. and Royall, R. M. (1999). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley.

Received March 9, 2006; accepted July 26, 2006.

Gordon G. Bechtel
University of Florida and Florida Research Institute
P.O. Box 117155
Gainesville, Florida 32611-7155, USA
bechtel@ufl.edu