

# A GENERAL FRAMEWORK FOR ESTIMATING A REGRESSION MODEL WITH A NATURAL EXTENSION TO THE ANALYSIS OF SURVEY DATA

Phillip S. Kott<sup>1</sup>

## ABSTRACT

In statistics we often estimate models that we know are not true representations of reality. A sensible strategy in such a situation is to loosen the assumptions underpinning the model. This paper relaxes the strong assumption that the error term in a not-necessarily-linear regression model has mean zero given any set of realized values for the independent variables. In its place is the much weaker assumption that the error term has mean zero unconditionally and is simply uncorrelated with the independent variables. This general regression framework is then extended to the estimation of a regression model with survey data. In so doing, a unified approach to estimating a regression model emerges.

KEY WORDS: Asymptotic; Cluster; Error term; Primary sampling unit; Selection probability; Stratum.

## RÉSUMÉ

En statistique, nous estimons souvent des modèles tout en sachant qu'ils ne représentent pas vraiment la réalité. Une stratégie sensée lors d'une telle situation est de relâcher les hypothèses soutenant le modèle. Dans cet article, on relâche la forte hypothèse voulant que le terme d'erreur dans un modèle de régression - qui n'est pas nécessairement linéaire - ait zéro pour moyenne peu importe l'ensemble de valeurs obtenues pour les variables indépendantes. Au lieu, nous émettons l'hypothèse beaucoup plus faible que le terme d'erreur a inconditionnellement une moyenne égale à zéro et qu'il n'est pas corrélé avec les variables indépendantes. Ce cadre général de travail pour la régression est ensuite élargi à l'estimation d'un modèle de régression avec des données provenant d'enquêtes. Ce faisant, une approche unifiée pour estimer un modèle de régression est obtenue.

MOTS CLÉS: Asymptotique; grappe; terme d'erreur; unité primaire d'échantillonnage; probabilité de sélection; strate.

## 1. INTRODUCTION

"All models are wrong but some are useful" is a truism about statistical modeling generally attributed to George Box. Knowing it is likely that one's model is wrong, a sensible strategy in many situations is to loosen the assumptions underpinning the model. This note relaxes the assumption that the error term in a not-necessarily-linear regression model has mean zero conditioned on the realized values of the independent variables.

The framework developed here is not really new. It builds on ideas discussed by White (1980) and Hansen (1982) in the econometrics literature and Fuller (1975) in the survey sampling literature. It is very closely linked to the treatment of linear regression found in

Magee (1997) and differs in many respects from my previous efforts in this area (for example, Kott, 1991).

For those interested in such things, many of the missing proofs, especially those concerning non-linear regression, can be obtained by modifying analogous results in Binder (1983). More important to this author are the conceptual differences between the approach taken here and that found in Binder.

Section 2 lays out the basic framework. Section 3 provides a simple example of how a zero-meanned error term can be uncorrelated with an independent variable but have a mean other than zero when conditioned on it. Section 4 contains some asymptotic (both large population and large sample) theory and begins the treatment of variance estimation. Section 5 discusses alternatives to "simple linear regression through the

---

<sup>1</sup> Phillip S. Kott, U.S. National Agricultural Statistics Service, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030, USA, pkott@nass.usda.gov

origin" in light of the framework of the previous sections. Section 6 introduces complex survey sampling into the analysis, with Section 7 providing the necessary asymptotics for this extension. Section 8 develops variance estimation with a complex sample. Finally, Section 9 summarizes the results and places them into a broader context.

## 2. THE FRAMEWORK

Suppose we are interested in estimating the following stochastic model which describes a relationship among variables in a population:

$$y_i = F(\mathbf{z}_i; \boldsymbol{\beta}) + \varepsilon_i, \quad (1)$$

where  $F$  has a known form,  $i (= 1, \dots, M)$  denotes an element of the population,  $\mathbf{z}_i$  is a  $(p-1)$ -component vector of variable values associated with element  $i$ ,  $\boldsymbol{\beta}$  is an unknown  $p$ -component column vector, and  $\varepsilon_i$  is a random variable with mean zero that is uncorrelated with each component of  $\mathbf{z}_i$ . In the special case of linear regression,  $F(\mathbf{z}_i; \boldsymbol{\beta}) = \beta_1 + \sum_{1 < k \leq p} \beta_k z_{ik}$ . Other popular forms for  $F$  are  $[1 + \exp(\beta_1 + \sum_{1 < k \leq p} \beta_k z_{ik})]^{-1}$  and  $\exp\{\beta_1 + \sum_{1 < k \leq p} \beta_k \log(z_{ik})\}$ .

Our assumption about the "error term"  $\varepsilon_i$ , in equation (1) can be rendered  $E(\mathbf{x}_i' \varepsilon_i) = \mathbf{0}_p$ , where  $\mathbf{x}_i = (1, \mathbf{z}_i')$ . This is much weaker – and thus more general – than the conventional assumption,  $E(\varepsilon_i | \mathbf{x}_i) = 0$  for every possible  $\mathbf{x}_i$ . The latter implies that  $\varepsilon_i$  is not only uncorrelated with the components of  $\mathbf{x}_i$  but also with any function of the components of  $\mathbf{x}_i$ .

If every member of the population is an equally likely realization of the model in equation (1), then  $E[\sum \mathbf{x}_i' (y_i - F(\mathbf{z}_i; \boldsymbol{\beta}))] = \mathbf{0}_p$ . This suggest we estimate  $\boldsymbol{\beta}$  with the vector  $\mathbf{b}$  that satisfies

$$\sum \mathbf{x}_i' y_i = \sum \mathbf{x}_i' F(\mathbf{z}_i; \mathbf{b}). \quad (2)$$

Since this is essentially a  $p$  equation system with  $p$  unknowns, a unique solution exists under mild conditions. In the case of a linear regression model, equation (1) leads us to the ordinary least squares (OLS) solution  $\mathbf{b} = (\sum \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum \mathbf{x}_i' y_i$ .

## 3. AN EXAMPLE

The following example shows why the assumption  $E(\mathbf{x}_i' \varepsilon_i) = \mathbf{0}_p$  is a useful alternative to the standard assumption  $E(\varepsilon_i | \mathbf{x}_i) = 0$ . Suppose we have a population in which the relationship  $y_i = z_i^\gamma$  for  $\gamma \neq 1$  is strictly satisfied. We do not know this, however, and try to set  $F$  in equation (1) equal to  $\beta_1 + \beta_2 z_i$ .

The OLS estimates for  $\beta_2$  and  $\beta_1$  are  $b_2 = (\sum z_i^{\gamma+1} - \sum z_i^\gamma \sum z_i / M) / (\sum z_i^2 - [\sum z_i]^2 / M)$ , and  $b_1 = \sum (z_i^\gamma - b_2 z_i) / M$ , respectively. We make the relatively mild assumption that the series  $\sum z_i^\alpha / M$  converges to a constant, say  $z^{(\alpha)}$ , as  $M$  grows arbitrarily large, where  $\alpha$  can have any of the following values: 1, 2,  $\gamma$ , or  $\gamma+1$ . Under this assumption,  $b_2$  converges to  $\beta_2 = (z^{(\gamma+1)} - z^{(\gamma)} z^{(1)}) / (z^{(2)} - [z^{(1)}]^2)$ , and  $b_1$  to  $\beta_1 = z^{(\gamma)} - \beta_2 z^{(1)}$ .

It is now easy to see that  $\varepsilon_i = z_i^\gamma - (\beta_1 + \beta_2 z_i) = (z_i^\gamma - z^{(\gamma)}) - \{(z^{(\gamma+1)} - z^{(\gamma)} z^{(1)}) / (z^{(2)} - [z^{(1)}]^2)\} (z_i - z^{(1)})$  has mean zero and is uncorrelated with  $z_i$  (at least in the sense that  $\text{plim}(\sum \varepsilon_i / M) = \text{plim}(\sum z_i \varepsilon_i / M) = 0$ ). In contrast to this, the value of  $E(\varepsilon_i | z_i)$  is not equal to zero for all  $z_i$ . This is because although  $\varepsilon_i$  is uncorrelated with  $z_i$  it is clearly a function of  $z_i$ .

The example shows that the assumption  $E(\mathbf{x}_i' \varepsilon_i) = \mathbf{0}_p$  allows a flexibility in model construction that is unavailable with  $E(\varepsilon_i | \mathbf{x}_i) = 0$ . Since reality very seldom fits a postulated model, this flexibility is fortuitous. In our example, when  $E(\varepsilon_i | \mathbf{x}_i) = 0$  is assumed, the model  $y_i = \beta_1 + \beta_2 z_i + \varepsilon_i$  is simply wrong, and its parameters *cannot* be estimated. When  $E(\mathbf{x}_i' \varepsilon_i) = \mathbf{0}_p$  is assumed, however, the parameters of the model *can* be estimated. Many will argue that we should not estimate parameters for "wrong" models, but as Box noted aren't all models wrong?

## 4. SOME THEORY

Suppose the model in equation (1) holds, and equation (2) has a solution. Call it  $\mathbf{b}^*$ . For  $\mathbf{b}^*$  to be a consistent estimator for  $\boldsymbol{\beta}$ , a number of asymptotic conditions must be satisfied. It is sufficient that

$$\lim_{M \rightarrow \infty} (\sum \mathbf{x}_i' F(\mathbf{z}_i; \mathbf{b}) / M) = G(\mathbf{b}), \quad (3.1)$$

for every possible  $\mathbf{b}$ ,

$$\text{plim}_{M \rightarrow \infty} (\sum \mathbf{x}_i' \varepsilon_i / \sqrt{M}) = \mathbf{d}, \quad (3.2)$$

for some bounded vector  $\mathbf{d}$  and  $G(\mathbf{b})$  be twice differentiable in the neighborhood of  $\mathbf{b}^*$ . Under these conditions, one can show that  $\mathbf{b}^* - \boldsymbol{\beta} = \mathbf{O}_p(\sqrt{1/M})$ .

Without loss of generality, we now let  $\mathbf{b}$  denote the solution to equation (2). We can rewrite (2) as  $[\sum \mathbf{x}_i' y_i - \sum \mathbf{x}_i' F(\mathbf{z}_i; \mathbf{b})] / M = 0$ , which means that  $[\sum \mathbf{x}_i' \varepsilon_i + \sum \mathbf{x}_i' \{F(\mathbf{z}_i; \boldsymbol{\beta}) - F(\mathbf{z}_i; \mathbf{b})\}] / M = 0$ . Assuming  $F$  is twice differentiable around  $\boldsymbol{\beta}$ , and letting  $\mathbf{f}_i$  be the row vector of partial derivatives of  $F(\mathbf{z}_i; \boldsymbol{\beta})$  taken with respect to the components of  $\boldsymbol{\beta}$  and evaluated at  $(\mathbf{z}_i; \boldsymbol{\beta})$ , we have  $[\sum \mathbf{x}_i' \varepsilon_i - \sum \mathbf{x}_i' \mathbf{f}_i (\mathbf{b} - \boldsymbol{\beta})] / M \approx 0$ . This in turn implies  $\mathbf{b} \approx \boldsymbol{\beta} + (\sum$

$\mathbf{x}_i' \mathbf{f}_i)^{-1} \sum \mathbf{x}_i' \varepsilon_i$ , which gives us a hook into estimating the variance of  $\mathbf{b}$  as long as  $\sum \mathbf{x}_i' \mathbf{f}_i / M$  is invertible.

Let  $\mathbf{u}_i = \mathbf{x}_i' \varepsilon_i$ , and suppose the population can be grouped into  $J$  mutually exclusive clusters, denoted  $C(1), \dots, C(J)$ , such that  $E(\mathbf{u}_i \mathbf{u}_k')$  is non-negative definite when  $i$  and  $k$  are in the same cluster and equal to  $\mathbf{0}_{p \times p}$  otherwise. In many practical situations, the  $M$  elements in the population will serve as the  $J$  clusters. In others, there will be a clear need to collect in clusters elements whose error terms can not be assumed uncorrelated, as we shall see in the following section.

If  $J/M$  converges to a positive constant as  $M$  grows arbitrarily large, then under mild conditions the "sandwich estimator,"

$$\mathbf{V} = (\sum^M \mathbf{x}_i' \mathbf{f}_i)^{-1} \sum^J (\sum_{i \in C(j)} \mathbf{r}_i \mathbf{r}_i') (\sum^M \mathbf{x}_i' \mathbf{f}_i)^{-1}, \quad (4)$$

where  $\mathbf{r}_i = \mathbf{x}_i'(y_i - F(\mathbf{z}_i; \mathbf{b})) \approx \mathbf{u}_i$ , is a consistent estimator for the variance of  $\mathbf{b}$  (a positive definite variance estimator,  $\mathbf{V}$ , for a consistent estimator like  $\mathbf{b}$  is itself consistent when the relative mean squared error of  $\mathbf{V}\lambda$  is asymptotically zero for all non-trivial  $p$ -vectors  $\lambda$ ). Since  $\mathbf{b}$  is consistent its variance and mean squared error are asymptotically indistinguishable.

## 5. ALRNATIVES TO SIMPLE LINEAR REGRESSION THROUGH THE ORIGIN

Although equation (2) can provide a solution for many models that do not fit population data all that well (in that  $E(\varepsilon_i | \mathbf{x}_i) \neq 0$ ), it may not work for one of the most popular models, namely, simple linear regression through the origin:  $y_i = \beta z_i + \varepsilon_i$ . The problem here is that there are two equations we want satisfied,  $\sum^M (y_i - \beta z_i) = 0$  and  $\sum^M (y_i - \beta z_i) z_i = 0$ , but there is only one unknown parameter,  $\beta$ .

The solution is to reparameterize the problem as  $y_i/z_i = \beta + \varepsilon_i/z_i$  or  $y_i^* = \beta + \varepsilon_i^*$ . Now there is only one equation we need be satisfied:  $\sum^M (y_i^* - \beta) = 0$ . Its solution is  $\mathbf{b} = \sum^M y_i^* / M = \sum^M (y_i/z_i) / M$ .

There is an alternative estimator for  $\beta$  that may be more appropriate in certain applications. An example will explain why. Consider a population of  $N$  farms, where farm  $i$  has  $z_i$  acres of farmland on which  $y_i$  acres of corn have been planted. Viewed as a population of farms,  $b_F = \sum^N (y_i/z_i) / N$  is a consistent estimator of  $\beta$ . Viewed as a population of  $M = \sum^N z_i$  acres, however, we get a different estimator,  $b_A = (\sum^N \sum_{k \in i} y_{ik}) / (\sum^N \sum_{k \in i} 1) = \sum^N y_i / \sum^N z_i$ , where  $y_{ik}$  is the fraction of corn acres on farmland acre  $k$  of farm  $i$ . Not only are the two estimators usually different, but if  $E(\varepsilon_i | z_i) \neq 0$ , their target  $\beta$ -values can also be different.

The choice between  $b_F$  and  $b_A$  is not a statistical one.

The user of the statistical product needs to determine whether (s)he is more interested in what is happening on the average farm or what is happening on the average acre of farmland. This is a common problem when estimating ratios that is not limited to agriculture. When  $E(\varepsilon_i | z_i) \neq 0$ , the user must choose her target.

The farm/farmland example can provide a useful insight into variance estimation with equation (4). When farms are the designated population elements, it may be reasonable to treat the element error terms, the  $\varepsilon_i^*$ , as uncorrelated. Thus, each cluster in (4) consists of a single element. When farmland acres are the population elements, however, it makes sense to collect the  $z_i$  elements in farm  $i$  into a cluster,  $C(i)$ , so that the element errors, the  $\varepsilon_{ik} = y_{ik} - \beta$ , for different acres from the same farm are allowed to be correlated.

## 6. RANDOM SAMPLING

Solving equation (2) to derive an estimator  $\beta$  assumes that the  $M$  elements in the population are generated by a process which produces elements satisfying equation (1). Moreover, were the process allowed to continue, the two parts of equation (3) would likewise be satisfied.

Following Fuller (1975), we will treat the  $J$  clusters in the population as if they were a simple random sample from a putative infinite population, each of whose elements satisfy equation (1). Moreover, as the number of these clusters (and therefore  $M$ ) grows arbitrarily large equation (3) will be assumed to hold.

Sometimes we do not have access to information on all the variables in equation (1) for the entire population. Instead, a probability sample is drawn, and a complete set of variable values are collected only for the sample. We will concentrate here on a stratified, multi-stage sample and ignore the possibility of element or item (i.e., variable) nonresponse.

Suppose that, before sampling, the  $J$  clusters in the population are separated into  $H$  mutually exclusive strata ( $H$  may be 1). A probability sample of  $n_h$  clusters are selected within each stratum  $h$  without replacement (from now on, all samples are assumed to be drawn without replacement). The  $n = \sum n_h$  sampled clusters are called primary sampling units (PSU's). Probability samples of elements are drawn independently within each PSU. We allow the extreme possibilities that either all the elements in a PSU are drawn into the sample or that the PSU's are themselves elements. Let  $S$  denote the element sample and  $m$  be the size of  $S$ .

If  $E(\mathbf{u}_i | i \in S) = \mathbf{0}_p$ , then solving equation (2) — if possible — with the summations taken over the sample rather than the population would provide a consistent

estimator for  $\boldsymbol{\beta}$  under mild additional conditions (essentially replacing  $M$  in equation (3) by  $m$ ). The assumption that  $E(\mathbf{u}_i | i \in S) = \mathbf{0}_p$  effectively means that there is no information about  $y_i$  in the element selection probabilities not captured by  $F(\mathbf{x}_i; \boldsymbol{\beta})$ . What if that is not the case?

The solution will not surprise anyone familiar with randomization-based inference (see, for example, Binder (1983)). Let  $I_i = 1$  when  $i \in S$  and 0 otherwise. Furthermore, let  $\pi_i$  be the selection probability of element  $i$ .

Suppose the vector  $\mathbf{b}_\pi$  solves the equation:

$$\sum^M \mathbf{x}_i' y_i (I_i / \pi_i) = \sum^M \mathbf{x}_i' F(\mathbf{z}_i; \mathbf{b}) \mathbf{x}_i' y_i (I_i / \pi_i) \quad (2)$$

Observe that in this formulation although the summation is over the entire population, only the elements in the sample have non-zero values. If we assume that the variables and sample design are such that

$$\lim_{n \rightarrow \infty} (\sum [I_i / \pi_i] \mathbf{x}_i' F(\mathbf{z}_i; \mathbf{b}) / M) = G(\mathbf{b}) \quad (3.1)$$

for every possible  $\mathbf{b}$ ,

$$\text{plim}_{n \rightarrow \infty} (\sum [I_i / \pi_i] \mathbf{x}_i' \varepsilon_i / \sqrt{n}) = \mathbf{d}_\pi \quad (3.2)$$

for some bounded vector  $\mathbf{d}_\pi$ , then under mild conditions,  $\mathbf{b}_\pi - \boldsymbol{\beta} = \mathbf{O}_p(\sqrt{1/n})$ .

## 7. ASYMPTOTICS FOR A COMPLEX SAMPLE

In order to apply equation (3), we need to impose an asymptotic framework on the sample design. We do this by assuming a infinite sequence of samples and populations,  $\{S_v\}$  and  $\{P_v\}$ . Let  $m_v$  denote the number of elements in  $S_v$ ,  $M_v$  the analogous size of  $P_v$ ,  $n_v$  the number of PSU's in  $S_v$ ,  $J_v$  the number of clusters in  $P_v$ ,  $H_v$  the number of strata in both  $S_v$  and  $P_v$ , and  $n_{hv}$  the number of PSU's from stratum  $h$  in  $S_v$ .

As  $v$  grows arbitrarily large, so does  $n_v$ . The ratios,  $m_v / n_v$ ,  $J_v / n_v$ , and  $M_v / m_v$  all converge to positive constants. When  $H$  is small, it makes sense to assume an asymptotic framework in which  $H_v$  stays the same as  $v$  grows, and the  $n_{hv} / n_v$  converge to positive constants. Otherwise, the  $n_{hv}$  can be assumed stay the same while  $H_v / n_v$  converges to a positive constant. *It is important to realize that equation (3) with  $M_v$  replacing  $M$  is assumed to hold for each value of  $v$ .*

## 8. VARIANCE ESTIMATION FOR $\mathbf{b}_\pi$

If  $E(\mathbf{u}_i \mathbf{u}_k' | i, k \in S) = \mathbf{0}_{p \times p}$  when  $i$  and  $k$  are from different clusters (and non-negative definite otherwise), then under mild conditions we could estimate the variance of  $\mathbf{b}_\pi$  consistently with an analogue of  $\mathbf{V}$  in equation (4), namely,

$$\mathbf{V}' = \left( \sum^M [I_i / \pi_i] \mathbf{x}_i' \mathbf{f}_i \right)^{-1} \sum^J \left( \sum_{i \in C(j)} [I_i / \pi_i] \mathbf{r}_i \mathbf{r}_i' \right) \left( \sum^M [I_i / \pi_i] \mathbf{x}_i' \mathbf{f}_i \right)^{-1} \quad (5)$$

where  $c(j)$  denotes the element sample in PSU  $j$ , and  $\mathbf{r}_i$  is now  $\mathbf{x}_i'(y_i - F(\mathbf{z}_i; \mathbf{b}_\pi))$ . In some applications, however,  $E(\mathbf{u}_i \mathbf{u}_k' | i, k \in S)$  may not equal  $\mathbf{0}_{p \times p}$  when  $i$  and  $k$  from are different clusters. Before discussing a more general variance estimator than  $\mathbf{V}'$ , we first investigate the potential effect of stratification on estimation. Let  $T_i$  be an integer from 1 to  $H$  indicating which stratum contains  $i$ . Although it is often tempting to assume that  $E(\mathbf{u}_i | T_i) = \mathbf{0}_p$ , in some applications this equality is not very likely. Consider, for example, a population of third-graders, stratified by sex, in which test scores for a sample of students are being modeled as a function of race and ethnicity exclusively. Here, even though  $E(\mathbf{u}_i) = \mathbf{0}_p$ ,  $E(\mathbf{u}_i | T_i = 1)$ , with  $T_i = 1$  meaning that  $i$  is male, may not be  $\mathbf{0}_p$ .

Suppose we adopt the assumption  $E(\mathbf{u}_i | T_i) = \mathbf{t}_h$  for element  $i$  in stratum  $h$ . We could try to build a variance estimator for  $\mathbf{b}_\pi$  around this assumption, but difficulties under multi-stage sampling quickly arise. An alternative approach is to invoke the randomization-based properties of the estimator; that is, to treat the  $I_i$  as random variables rather than conditioning on the realized sample – the latter being the usual practice in most of statistics. Mathematically, we change the goal of variance estimation from estimating  $E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' | S]$  for every  $S$  to estimating  $E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})']$ .

Since we are allowing the possibility that  $\mathbf{t}_h$  varies across strata, we need assume that  $N_h / N$  stays constant as  $N$  and  $M$  grow arbitrarily large in equations (3) and (3'). This means that the fraction of the population clusters in each stratum does not change as the population grows arbitrarily large. If it did, there could be another component

of variance not captured by the variance estimator to be discussed below (See Korn and Graubard, 1994).

Let  $I_{i(h)} = 1$  when  $i$  is a sampled element in the  $v$ 'th PSU selected from stratum  $h$  (0, otherwise). Observe that the random variables  $\mathbf{g}(v; h) = \sum_{i \in C(j)} [I_{iv} / \pi_i] \mathbf{u}_i$  have a common expectation within each stratum;

namely,  $n_h^{-1} \sum^{Mh} \mathbf{u}_i$ . (Note:  $N_h$  and  $M_h$  are rendered  $N_h$  and  $M_h$  as superscripts).

If the PSU's are selected within stratum  $h$  using simple random (cluster) sampling, then each of the  $\delta_v = [n_h / M_h] \mathbf{g}(v; h)$  are asymptotically independent estimators of  $\text{plim}_{M \rightarrow \infty} (M_h^{-1} \sum^{Mh} \mathbf{u}_i)$ . This is because conducting two phases of simple random sampling within each stratum is equivalent to drawing a stratified, simple random sample. Selections from a simple random sample of an infinite population are asymptotically independent.

A consistent estimator of the variance of  $\sum^{nh} \delta_v / n_h$  as an estimator for  $\text{plim}_{M \rightarrow \infty} (M_h^{-1} \sum^{Mh} \mathbf{u}_i)$  under mild conditions is  $M_h^{-2} (n_h / [n_h - 1]) \{ \sum^{nh} \mathbf{g}(v; h) \mathbf{g}(v; h)' - \sum^{nh} \mathbf{g}(v; h) \sum^{nh} \mathbf{g}(v; h)' / n_h \}$ . Since,  $\mathbf{b}_\pi \approx \boldsymbol{\beta} + (\sum^M [I_i / \pi_i] \mathbf{x}_i' \mathbf{f}_i)^{-1} \sum^H \sum^{nh} \mathbf{g}(v; h)$ , it is now not difficult to see that the conventional, randomization-based variance estimator for  $\mathbf{b}_\pi$  (see Binder, 1983),

$$\mathbf{V}_{RB} = (\sum^M [I_i / \pi_i] \mathbf{x}_i' \mathbf{f}_i)^{-1} \sum^H \{ (n_h / [n_h - 1]) \sum^{nh} [(\sum_{i \in C(j)} [I_i / \pi_i] \mathbf{r}_i \mathbf{r}_i') - n_h^{-1} (\sum^{Nh} \sum_{i \in C(j)} [I_i / \pi_i] \mathbf{r}_i) (\sum^{Nh} \sum_{i \in C(j)} [I_i / \pi_i] \mathbf{r}_i)'] \} (\sum^M [I_i / \pi_i] \mathbf{x}_i' \mathbf{f}_i)^{-1}, \quad (6)$$

is consistent under mild conditions when PSU's are selected using stratified, simple random sampling. Moreover, in most practical situations,  $\mathbf{V}_{RB}$  will be reasonable – although not necessarily consistent – even when PSU's are selected with unequal probabilities within strata (see Kott, 1997).

Observe that when  $H = 1$  in equation (6),  $\mathbf{V}_{RB}$  becomes asymptotically indistinguishable from  $\mathbf{V}'$  in equation (5) because  $\sum^N \sum_{i \in C(j)} [I_i / \pi_i] \mathbf{r}_i = \mathbf{0}_p$ . If  $E(\mathbf{u}_i \mathbf{u}_k' | i, k \in S) = \mathbf{0}_{p \times p}$  when  $i$  and  $k$  are from different clusters is satisfied, then  $\mathbf{V}_{RB}$  can easily be shown to be consistent given any sample. The same would be true for  $\mathbf{V}'$ , which would also tend to be the more efficient variance estimator.

## 9. DISCUSSION

We have developed a framework for estimating a not-necessarily-linear regression model, equation (1), in a rather general setting where the set of observations under analysis is either a cluster sample from a conceptual infinite population or a stratified, multi-stage, subsample of such a sample.

What sets this framework apart from the conventional is the avoidance of two common assumptions about the zero-meaned error term,  $\varepsilon_i$ , which can translated mathematically into assumptions about the product  $\mathbf{u}_i = \mathbf{x}_i' \varepsilon_i$ . The first is  $E(\mathbf{u}_i | \mathbf{x}_i) = \mathbf{0}_p$ . The second is  $E(\mathbf{u}_i | i \in S) = \mathbf{0}_p$ , which is equivalent to  $E(\mathbf{u}_i | I_i) = \mathbf{0}_p$  (because  $E(\mathbf{u}_i | i \in S) = E(\mathbf{u}_i | I_i = 1) = \mathbf{0}_p$ , while that and  $E(\mathbf{u}_i) = E(\mathbf{u}_i | I_i$

$= 1) \pi_i + E(\mathbf{u}_i | I_i = 0)(1 - \pi_i) = \mathbf{0}_p$  implies  $E(\mathbf{u}_i | I_i = 0) = \mathbf{0}_p$ ). In place of  $E(\mathbf{u}_i | \mathbf{x}_i) = \mathbf{0}_p$ , the framework developed here substitutes the weaker assumption  $E(\mathbf{u}_i) = \mathbf{0}_p$ . In place of  $E(\mathbf{u}_i | I_i) = \mathbf{0}_p$ , it substitutes  $E(\mathbf{u}_i | 1/\pi_i | I_i) = \mathbf{0}_p$ . The two strong assumptions ( $E(\mathbf{u}_i | \mathbf{x}_i) = \mathbf{0}_p$  and  $E(\mathbf{u}_i | I_i) = \mathbf{0}_p$ ) are similar, especially when  $\pi_i$ , the selection probability of element  $i$ , is a function of  $\mathbf{x}_i$ , but they are not the same thing. Magee (1997) discusses estimating a linear regression model when the first strong assumption is made but not the second.

Some may argue that regression has no meaning when  $E(\mathbf{u}_i | \mathbf{x}_i) \neq \mathbf{0}_p$ , while others take the same position when  $E(\mathbf{u}_i | I_i) \neq \mathbf{0}_p$ . Granted, the model in equation (1) is more informative when both strong assumptions hold. Nevertheless, the model still reveals something about the relationship between the variables in the population when one or both of the assumption fails.

In some situations it will be reasonable to assume that  $E(\mathbf{u}_i | \mathbf{x}_i) = \mathbf{0}_p$ . As a result, more efficient estimators for  $\boldsymbol{\beta}$  than the ones discussed here may be available. Magee addresses this issue in the linear case both when  $E(\mathbf{u}_i | I_i) = \mathbf{0}_p$  applies and when it fails. Our prime concern here, however, has not been efficiency but robustness to inevitable model failure.

One way to express the weak assumptions invoked here in compact form is with  $E(\mathbf{u}_i) = E(\mathbf{u}_i | \pi_i) = \mathbf{0}_p$ . If in addition,  $E[(\mathbf{u}_i | I_i / \pi_i)(\mathbf{u}_k | I_k / \pi_k)] = \mathbf{0}_{p \times p}$  for  $i$  and  $k$  from different PSU's (and non-negative definite otherwise), then variance estimation for  $\mathbf{b}_\pi$ , the solution to equation (2'), would be relatively simple:  $\mathbf{V}'$  from equation (5) is consistent under mild conditions. When  $E[(\mathbf{u}_i | I_i / \pi_i)(\mathbf{u}_k | I_k / \pi_k)] \neq \mathbf{0}_{p \times p}$  for  $i$  and  $k$  from different PSU's, then  $\mathbf{V}_{RB}$  from equation (6) is usually a reasonable variance estimator as we saw in the previous section.

The equality  $E[(\mathbf{u}_i | I_i / \pi_i)(\mathbf{u}_k | I_k / \pi_k)] = \mathbf{0}_{p \times p}$  for  $i \neq k$  holds when elements are selected via Poisson sampling. This is a situation investigated by Magee (1997), which effectively advocates  $\mathbf{V}'$  as the variance estimator in this context. When faced with a stratified sample, Magee essentially treats the strata containing sampled elements as a finite set of realizations from a infinite set of possibilities. In our framework, which better reflects reality, the set of strata containing sampled elements is fixed and exhausts the population even as the population grows arbitrarily large.

Binder's (1983) framework for estimating finite population parameters using complex survey data begins with a formula very much like equation (2). That approach treats the  $\mathbf{b}$  that solves this equation for the finite population as the target of estimation. Unlike here, there is no model in Binder's framework.

Since equation (2) comes directly from the assumption that  $\varepsilon_i$  is uncorrelated with  $\mathbf{x}_i$  and not from a

likelihood function, there is little reason to expect the estimators discussed here to match those derived from a likelihood-based analysis. It is serendipity that such a match appears in the case for logistic regression, where  $y_i$  is either zero or one,  $F(\mathbf{x}_i; \boldsymbol{\beta}) = [1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})]^{-1}$ , and  $\mathbf{x}_i = (1, \mathbf{z}_i)$ . In the conventional logistic regression model, one begins with the assumption  $\Pr(y_i = 1 | \mathbf{x}_i) = [1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})]^{-1}$ . No such assumption is required to set  $F(\mathbf{x}_i; \boldsymbol{\beta}) = [1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})]^{-1}$  in equation (1), however. This form for  $F$  need be no more than a convenient way to summarize the relationship between a vector of independent variables and a dichotomous dependent variable.

Not every non-linear regression model can be estimated directly or uniquely using the framework developed here. Section 5 addresses simple linear regression through the origin, where equation (2) translates into a system of two equations with only one unknown. In this situation, two potential reparameterizations are offered both of which lead to solutions within the framework.

Now consider the model:  $y_i = \alpha z_i + \beta z_i^2 + \varepsilon_i$ . Observe that since  $z_i$  is the lone independent variable, equation (2) becomes a system of two equations with three unknowns. One possible solution here would be to require  $\varepsilon_i$  to be uncorrelated with both  $z_i$  and  $z_i^2$ . The added requirement is appealing but arbitrary. Moreover, such a simply remedy may not always be available.

There can be a degree of arbitrariness even in relatively straight-forward problems. Consider the model:  $y_i = \alpha z_i^\beta + \varepsilon_i$ . Equation (2) now becomes a system with two equations and two unknowns and so has a unique set of solutions under mild conditions. Observe, however, that if we reparameterize the model as  $y_i^* = \alpha^* z_i^{\beta^*} + \varepsilon_i^*$ , where  $y_i^* = y_i / z_i$ , and  $\varepsilon_i^* = \varepsilon_i / z_i$ , we usually get a different unique set of solutions. This set may not be asymptotically identical to  $\alpha$  and  $\beta-1$  unless  $E(\varepsilon_i | z_i) = 0$ , an assumption we have been trying to avoid.

As with our two rival reparameterizations of simple linear regression through the origin, the choice between the two non-linear forms above can not be decided on statistical grounds. It is up to the user of the statistical results to decide which version of the dependent variable is of more interest.

Finally, it should be noted that the method-of-moments approach to the estimation of regression parameters, introduced by Hansen (1982), routinely

treats situations in which there are more equations than unknowns. In the methods-of-moments literature, all the equations are assumed to be correct. Here, that assumption is made only when the number of equations equals the number of unknown parameters. This is an important distinction. Nevertheless, it would not be unfair to call the implicit solution for the parameter vector in equation (2) or (2'): an application of method of moments in the trivial case.

## REFERENCES

- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya*, Ser. C, 37, 117-132.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments, *Econometrica*, 50, 1029-1054.
- Korn, E.L. and Graubard, B.I. (1994). Variance estimation for superpopulation parameters: should one use with replacement estimators? *American Statistical Association, Proceedings of the Survey Research Methods Section*, 124-131.
- Kott, P.S. (1991). A model-based look at linear regression with survey data, *American Statistician*, 45, 107-112.
- Kott, P.S. (1997). A note on the infinite-population randomization-based approach to the analysis of survey data. Submitted to *Journal of Official Statistics*.
- Magee, L. (1997). Improving survey-weighted least squares regression. Forthcoming *Journal of the Royal Statistical Society B*.
- White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review*, 21, 149-170.