REGRESSION ANALYSIS IN SAMPLE SURVEYS*

H. S. Konijn

University of Sydney and Cowles Foundation, Yale University

Models are proposed as a basis for unbiased and biased estimation from sample survey data of a regression relation. Unbiased estimates of the variances and covariances of estimates of the regression coefficients are obtained for certain cases but do not appear to exist for all cases. It is suggested that past surveys, conducted mostly to obtain efficiently estimates of means and totals, be analyzed to yield estimates of regression relations, to obtain information needed in designing future surveys, and to test the adequacy of the proposed models.

1. INTRODUCTION

In a great many surveys the observations have not been obtained by unrestricted random sampling, but stratification and/or cluster sampling has been employed. The usual estimators of the coefficients of a regression relation for the population as a whole (or some broad section of it) which are of the form of \( a' \) and \( b' \) defined in the sequel, are shown to be unbiased when the regression slopes are the same in all strata or clusters.

However, we show that even small deviations from the condition of equal slopes may lead to large biases. (In this connection it should be noted that very often the number of observations within the penultimate sampling units is too small for detecting small variations in slope with any reasonable degree of certainty.) Thus we are led to define another pair of estimators, \( a \) and \( b \), of the regression constants which are unbiased even when the regression slopes differ within penultimate sampling units. To permit a full study of \( a, b, a' \) and \( b' \), it is necessary to formulate a model. We propose two alternative models, which differ somewhat in their implications.

A procedure occasionally used in surveys to estimate a total or a mean, known as the method of regression estimation, should not be confused with the subject of the present paper.

For the reader's convenience we assemble the different notations used and their definition in an appendix.

2. THE STRUCTURE OF THE SURVEY AND THE PROBLEM

Consider the following situation. The surveyed population consists of \( M \) objects or "individuals" belonging to \( N \) classes which—purely for reasons of convenience and efficiency of sampling—are sampled separately. Let class number \( v \) contain \( M_v \) individuals. The sample may contain \( n \) out of the \( N \) classes and, if class \( v \) is included, \( m_v \) from class \( v \). When all classes are included in the sample, they are referred to as strata, otherwise as clusters. The method outlined below can also be extended to combinations of stratified and cluster sampling; these occur frequently in practice.

Each individual has two observable characteristics \( y \) and \( x \); we wish to make inferences about a regression of \( y \) on \( x \) (our method can be generalized to

multiple regression without difficulties). Let the conditional expectation of the 
y value for an individual drawn at random from the population, given his x 
value, be \( \alpha + \beta x \) and denote the residual \( y - (\alpha + \beta x) \) by \( z \), with the assumed 
mean of \( z \) given \( x \) equal to zero. We shall investigate estimates of \( \alpha \) and \( \beta \) and 
their variances.

3. TWO MODELS

We shall specify two different models which both entail the above regression 
relation and allow us to compute point estimates and their variances.\(^1\)

3.1 THE FIRST MODEL

In a first model we suppose that the population of \( M \) individuals itself con-
stitutes a proportionate stratified sample from a (conceptual) infinitely large 
population of individuals with similar behavior. This makes the residual \( z \) an 
ordinary random variable, in the sense of a drawing from an infinite popu-
lation, uncorrelated with each \( x \) and with zero mean. The drawings from 
the infinite population are supposed to have been made independently in the di-

erent strata.

We now suppose that in each stratum of the infinite population a regression 
relation holds of the same form and with the same properties as the relation 
which we wish to estimate; i.e., that the conditional expectation of \( y \), given 
that it is drawn from class \( v \), and given the value of \( x \), is \( \alpha_v + \beta_v x \). Concretely 
this means that the average value of \( y \) for an individual in the subclass of 
the infinite population whose individuals belong to class \( v \) and who have a given 
\( x \) value \( x_0 \)—in other words that the mean of \( y \) for an individual selected at 
random from the subclass \( v \) in the finite population with \( x \) value \( x_0 \)—is, if that 
subclass is not empty, \( \alpha_v + \beta_v x_0 \). As usual it follows that the residual \( z \) has a 
mean, given \( x \), of zero.\(^2\)

In the computation of variances of our estimates we also assume that the 
mean \( \sigma_x^2 \) of \( z^2 \) for individuals selected at random from \( v \) with given \( x \) value is 
positive and independent of this \( x \) value (from footnote 2 we see that the 
mean of the product of residuals for pairs of individuals selected at random in 
the subpopulation with given \( x \) values is zero).

3.2 AN ALTERNATIVE MODEL

As a second model assume that for any given individual in the finite popu-
lation the conditional mean of \( y \) given \( x \) is linear in \( x \), and that the deviation \( z \) cor-
responding to a given individual and for a fixed value of \( x \) represents one 
realization from among a class of potential fluctuations about this individual's

\(^1\) In considering other possible models, one should note that it would be quite unreasonable to postulate that 
in the finite population a linear relation could be defined such that the deviations from it for any set of individuals 
with the same \( x \) value average out to exactly zero. Finite populations that are fully enumerated rarely if ever 
satisfy this condition.

\(^2\) We may also conclude that if we draw a sample of several individuals, the average value of functions of the 
residual for the first individual, in the subclass of class \( x \) for which the first individual has \( x \) value \( x_0 \) and the other 
individuals certain specified \( x \) or \( y \) values, does not depend on the \( x \) or \( y \) values other than \( x_0 \). This is because in the 
infinite population, the draws are independent so that a restriction on the values of \( x \) or \( y \) for, say, the second indi-

dvidual drawn does not restrict the possible \( y \) values of the first individual in the subpopulation with \( x \) values \( x_0 \) 
(unless \( y \) is a constant). This property is used in the derivation of the variance. For instance, the mean of the product 
of the \( z \) for two individuals with specified \( x \) values, \( \overline{z} \) (for \( z \), \( z_0 \)) = \( \overline{z} \) (for \( x_0 \), \( x_0 \)) = 0 since 
\( \overline{z} \) (for \( z \) \( x_0 \)) = \( \overline{z} \) (for \( x_0 \), \( x_0 \)) = \( \overline{z} \) (for \( x \), \( x \)) = 0.
conditional mean $y$ value for any fixed $x$. Such fluctuations would come about hypothetically by an indefinite repetition of the observed situation for that individual under essentially identical circumstances.

But in such a model neither the regression nor the conditional variance of $x$ could be estimated on the basis of the sample, since for each individual only one $x$ and $y$ observation is obtained, unless one has more than one observation per individual or unless some relation is assumed between the regression coefficients for the different individuals and between their conditional variances.

Thus one could assume that the regression coefficients and the conditional variances are positive and the same for all individuals in class $v$ and independent of $z$. Concretely this would mean that potential fluctuations in individual behavior in class $v$ are similar for all individuals in that class.

If, moreover, one specifies the constancy of the conditional covariances, say $r_{uv}$ for pairs of individuals in class $v$, and the vanishing of these covariances for individuals in different classes, all the results of the paper continue to apply if the $r_{uv}$ are all zero. Even if they are not, we shall see in section 7 that most results remain valid.

### 3.3 Additional Remarks on Both Models

For both models we see that the regression relation of section 2 is satisfied for an individual selected at random from that part of the entire population in which individuals have $x$ value $x_0$ with

$$a = \sum M_v \alpha_v / M, \quad \beta = \sum M_v \beta_v / M.$$ 

To avoid trivialities, we assume that $N$ and each $M_v$ exceed unity, and that we do not have both $\alpha_1 = \cdots = \alpha_N$ and $\beta_1 = \cdots = \beta_N$. Note that if the $\beta_v$ differ very much from each other, as compared with the $\sigma_0^2$, there is little point in estimating the overall relation of section 2, since the variance of the estimate of $\beta$ will be very large.

If the classes have ranges for $x$ which overlap only a little or not at all, it is not appropriate to adopt either of the above models, unless there is external evidence that the regression line within the classes may be extensively extrapolated. Without such evidence it would be better to adopt a model leading to an overall relation which is piecewise linear or nonlinear in other ways.

In the following we shall concentrate on the first model, but give a brief discussion for the second model in section 7.

### 4. The Selection Procedure

Denote the $n$ selected clusters by their serial numbers $v_1, \cdots, v_n$ (the $v_i$ need not be all different) in the set of $N$ classes, in the order in which they are drawn. That is, the $i$th selected cluster is cluster number $v_i$; if there is no likelihood of confusion we shall also denote this by cluster $i$ as is usual in the
sampling survey literature. The clusters are selected so that the probability of selecting the sequence \( r_1, \ldots, r_n \) is (for each such sequence) a known number
\[
\frac{1}{p_{r_1} \cdots p_{r_n}};
\]
see, e.g., [1].

In the \( i \)th selected cluster, a simple random sample \( R_i \) of \( m_i \) (\( > 1 \)) individuals is taken independently of the previous sampling process, giving rise to pairs of observations
\[
(x_{i1}, y_{i1}), \ldots, (x_{im_i}, y_{im_i}).
\]
We assume that, for each \( i \), not all of the numbers \( x_{i1}, \ldots, x_{im_i} \) coincide (or rather that the probability of coincidence is zero). The selection of the \( n \) simple random samples takes place independently in each of the chosen clusters.

The probabilities
\[
\frac{1}{p_{r_1} \cdots p_{r_n}}
\]
and the \( m_i \) are usually determined either by external or administrative exigencies or by the desire to keep down the variance of certain estimated totals, averages or ratios, rather than to obtain precise estimates of the regression coefficients. We shall therefore not inquire into their best choice.

In all that follows \( \nu \) will run over the integers 1 to \( N \), \( i \) over the integers 1 to \( n \); and, for fixed \( i, j \), \( \nu \) will run over the \( m_i \) individuals in \( R_i \). \( \Sigma \) will indicate the sum over \( \nu \) or \( i \), \( \sum \) summation over \( j \).

Let \( p^\nu_i \) be the probability of including cluster \( \nu \) as the \( i \)th selected cluster. For \( i < \nu \) let \( p^\nu_i \) be the probability of including cluster \( \nu \) on occasion \( i \) and cluster \( \nu' \) on occasion \( \nu' \); for \( i > \nu \) define it by \( p_{\nu'}^\nu \). Evidently, \( \Sigma p^\nu_i = 1 \) and \( \Sigma \nu p^\nu_i = p^\nu_i \). We denote by \( p_{\nu} \) the probability of including cluster number \( \nu \) and by \( p_{\nu,\nu'} \) the probability of including cluster number \( \nu \) and number \( \nu' \).

Clusters may be selected with or without replacement.

We shall first discuss the case of sampling of clusters without replacement.

5. SELECTION OF CLUSTERS WITHOUT REPLACEMENT

In this case
\[
\begin{align*}
p^\nu_i &= \sum_{i < \nu} \left( p^\nu_i + p^\nu_{i,\nu} \right), \\
p^\nu_i &= \sum_{i > \nu} \sum_{\nu' > \nu} \left( p_{\nu'}^{i,\nu} + p_{\nu'}^{i',\nu} \right),
\end{align*}
\]
and
\[
\begin{align*}
\sum_{\nu} p^\nu_i &= \sum_{i < \nu} \sum_{\nu'} p_{\nu'}^{i,\nu} + \sum_{i > \nu} \sum_{\nu' > \nu} p_{\nu'}^{i,\nu} = \sum_{i < \nu} \sum_{\nu > \nu} p^i_{\nu} + \sum_{i > \nu} \sum_{\nu'} p_{\nu'}^{i,\nu} \\
&= \sum_{i} \sum_{\nu > \nu} p^i_{\nu} + \sum_{\nu > \nu} \sum_{\nu'} p_{\nu'}^{i,\nu} = \sum_{\nu > \nu} \sum_{\nu > \nu} p_{\nu'}^{i,\nu} = \sum_{\nu > \nu} (n - 1) p^i_{\nu} \\
&= \sum_{i} (n - 1) p_{\nu} = (n - 1) p_{\nu},
\end{align*}
\]
so
\[
\begin{align*}
p^2 - p^2 &= \sum_{\nu' > \nu} \{ p_{\nu'} p^2_{\nu'} - p^2_{\nu'} \}.
\end{align*}
\]
Moreover, if, for any selected cluster \( i \), a constant \( u_{ri} \) is defined and independent of \( R_i \), then \( u_r \) is defined for all \( r \) and
\[
\sum u_{ri} = \sum p_r u_r,
\]
\[
\sum u_{ri}^2 = \sum p_r u_r^2 + \sum p_{rr'} p_{rr'} u_r u_{r'},
\]
\[
\sum u_{ri} = \sum p_r (1 - p_r) u_r + \sum \sum p_{rr'} (p_{rr'} - p_r p_r') u_r u_{r'}.
\]

5.1 Point Estimates of the Regression Coefficients

To obtain an unbiased and easily computed estimate of \( \beta \) we shall search for a linear combination
\[
b = \sum SB_{ij} y_{ij}
\]
of the \( y_{ij} \) which has mean \( \beta \). Given that cluster \( i \) is in the sample, the usual estimate
\[
\hat{\beta}_i = S y_{ij} (x_{ij} - \bar{x}_i) / S (x_{ij} - \bar{x}_i)^2
\]
is an unbiased estimate of the corresponding \( \beta_i \). So a weighted average of the \( \hat{\beta}_i \)
\[
b = \sum k_i \hat{\beta}_i
\]
will yield an unbiased estimate of \( \beta \), and
\[
B_{ij} = k_i (x_{ij} - \bar{x}_i) / S (x_{ij} - \bar{x}_i)^2.
\]

To find the suitable weights, note that
\[
\sum SB_{ij} y_{ij} = \sum SB_{ij} x_{ij} + \sum \alpha_s SB_{ij} + \sum \beta_s SB_{ij} x_{ij}
\]
\[
= \sum SB_{ij} x_{ij} + \sum \beta_i k_i
\]
will have mean \( \beta \) for all possible values of the \( \beta \)'s if and only if we take
\[
k_r = M_r / (p_r M).
\]

Note that the usual estimate of \( \beta \) is a statistic of the form
\[
\sum c_i S (x_{ij} - \bar{x}_i) y_{ij} / \sum c_i S (x_{ij} - \bar{x}_i)^2.
\]
This is a ratio estimate and cannot be unbiased in general unless it is known in advance that \( \beta_1 = \cdots = \beta_N \) and we take \( c_i = c_i \) but otherwise arbitrary. This leads us to define
\[
b' = \sum SB'_{ij} y_{ij}, \quad B'_i = c_i (x_{ij} - \bar{x}_i) / \sum c_i S (x_{ij} - \bar{x}_i)^2.
\]
The estimate \( b' \) may be useful if the range of the \( \beta_i \) is known and small compared with the variance of \( b' \). In the next section we derive the latter (5.2.2') and also see that the bias is a weighted average of the numbers \( \beta_i - \bar{\beta} \) with weights proportional to the conditional expected value given the included clusters of
\[
c_i (x_{ij} - \bar{x}_i)^2 / \sum c_i S (x_{ij} - \bar{x}_i)^2.
\]
Therefore, if these are correlated with the \( \beta_i \), as they may well be, the bias may be large even for small \( \beta_i - \bar{\beta} \).
We also estimate \( \alpha \) by a linear function
\[
a = \sum S A_{ij} y_{ij}
\]
of the \( y_{ij} \). The usual estimate of \( \alpha \) is
\[
\hat{\alpha}_i = \bar{y}_i - \hat{\beta} \bar{x}_i.
\]
Considerations similar to those above lead to
\[
A_q = k_i/m_i - B_{ij} \bar{x}_i,
\]
so that
\[
a = \sum k_i \hat{\alpha}_i.
\]
Again, a statistic of the form
\[
\sum k_i \hat{\alpha}_i - \sum c_i S(x_{ij} - \bar{x}_i) y_{ij}(\sum k_i \bar{z}_i)/\sum c_i S(x_{ij} - \bar{x}_i)^2
\]
can not be unbiased in general, even when the \( \alpha \)'s are known to be all equal. But this expression is unbiased when it is given that \( \beta_1 = \cdots = \beta_k \) and we take \( c_i = c \) and \( k_i' = k_i \). We therefore define
\[
a' = \sum S A_{ij} y_{ij}, \quad A_{ij}' = (k_i/m_i) - B_{ij} \sum k_i \bar{x}_i.
\]
Similar remarks apply to \( a' \) as to \( b' \).

5.2 THE VARIANCES OF THE ESTIMATED REGRESSION COEFFICIENTS AND THEIR COVARIANCE

We first obtain the variance \( \mathfrak{C}b \) of our estimate \( b \). Now
\[
b = \sum S B_{ij} \bar{x}_i + \sum \beta_i k_i,
\]
and the conditional mean given that the sequence 1, \( \cdots, n \) of clusters constitutes the sample of clusters, and given the values of \( x_i \), \( e_i \) \( \cdots \), \( n \) equals \( \sum \beta_i k_i \). Let us write a conditional mean given \( i \) as \( \mathfrak{c}_i \), a conditional mean given \( i \) and the values of the \( x \)'s as \( \mathfrak{c}_i x \), a conditional variance given the inclusion of clusters 1, \( \cdots, n \) as \( \mathfrak{V}_i \) and as \( \mathfrak{V}_i x \) if the values of the \( x \)'s are also given. Then
\[
\mathfrak{V}_i x b = \sum \mathfrak{c}_i (S B_{ij} \bar{x}_i)^2 = \sum \mathfrak{c}_i S B_{ij} \bar{x}_i^2
\]
\[
= \sum \mathfrak{c}_i S B_{ij} = \sum \sigma_{ij}^2 S(x_{ij} - \bar{x}_i)^2.
\]
Putting
\[
\mathfrak{e}_i^2 = \sigma_{ij}^2(S(x_{ij} - \bar{x}_i)^2)
\]
and recalling that the unconditional variance of a variate equals the mean of its conditional variance plus the variance of its conditional mean, we have successively
\[
\mathfrak{V}_i x b = \sum \mathfrak{c}_i \sigma_{ij}^2 \mathfrak{e}_i^2
\]
and
\[
\var U b = \varepsilon \sum \sigma_i^2 \frac{k_i^2}{p_i^2} + \varepsilon \left( \sum \beta_i \bar{k}_i \right)^2 - \beta^2 = \sum p_i \sigma_i^2 \frac{k_i^2}{p_i} + \sum p_i \left( 1 - p_i \right) \frac{k_i^2}{p_i} + \sum \sum \left( p_{i'} - p_i p_i \right) \beta_{i'} \beta_i \bar{k}_{i'} \bar{k}_i. \tag{5.2.2}
\]

In the case of stratified sampling the \( p_i \) and \( p_{i'} \) equal 1 and
\[
\var U b = M^2 \sum M_i^2 \sigma_i^2, \tag{5.2.3}
\]
in fact, it is easy to see that this is the only case in which \( \var U b \) is independent of \( \beta \).

Now
\[
b' = \sum SB_i \varepsilon_i \bar{y} + \sum (\beta_i - \beta) C_i + \beta, \quad \text{with} \quad C_i = S(B_i \bar{x}_i), \tag{5.2.4}
\]
so
\[
\varepsilon_{1 \ldots n} b' = \sum (\beta_i - \beta) C_i + \beta, \tag{5.2.5}
\]
and
\[
\varepsilon_{1 \ldots n} b' = \sum \sigma_i^2 S(B_i) \bar{y}^2.
\]
So
\[
\varepsilon_{1 \ldots n} b' = \sum \sigma_i^2 \varepsilon_{1 \ldots n} S(B_i \bar{y}) + \sum (\beta_i - \beta)^2 \varepsilon_{1 \ldots n} C_i,
\]
\[
\var U b' = \varepsilon \varepsilon_{1 \ldots n} b' + \varepsilon \left( \sum (\beta_i - \beta) \varepsilon_{1 \ldots n} C_i \right), \tag{5.2.2'}
\]
with the final terms vanishing when the \( \beta_i \) are equal. \( \var U b \) and \( \var U b' \) are not easily compared in general. If \( n = 1 \) and \( \beta_1 = \cdots = \beta_N \), the variances are equal when \( p_i = M_i / M \), but otherwise either one may exceed the other. The formulae show, however, that, even for a small range of the \( \beta \)'s, \( \var U b' \) may be substantially above the value for the case in which all \( \beta \)'s are equal.

When sampling with equal probabilities and \( N > 1 \),
\[
p_i = \frac{n}{N}, \quad p_{i'} = \frac{n(n - 1)}{N(N - 1)},
\]
and \( \var U b \) equals
\[
\frac{n}{N} \left\{ \sum \sigma_i^2 \frac{k_i^2}{p_i} + \frac{N - n}{N - 1} \sum (\beta_i \bar{k}_i - \beta \bar{k}_i)^2 \right\}, \tag{5.2.6}
\]
As
\[
a = \sum S A_i \varepsilon_i \bar{y} + \sum \alpha_k \bar{k}_i, \quad \varepsilon_{1 \ldots n} a = \sum \alpha_k \bar{k}_i,
\]
we have for the variance of \( a \) when sampling clusters without replacement:
\[
\var a = \varepsilon \sum \sigma_i^2 S A_i \frac{k_i^2}{p_i} + \varepsilon (\sum \alpha_k \bar{k}_i)^2 - \alpha^2
\]
\[
= \varepsilon \sum \sigma_i^2 \frac{k_i^2}{p_i} (m_i^{-1} + G_i) + \varepsilon (\sum \alpha_k \bar{k}_i)^2 - \alpha^2, \tag{5.2.7}
\]
where
\[ G_i^2 = \varepsilon \frac{\hat{x}_i^2}{S(x_{ij} - \bar{x})^2}; \]  
and for the covariance of \(a\) and \(b\):

\[ \varepsilon(a, b) = -\varepsilon \sum \hat{x}_i^2 H_i + \varepsilon(\sum \alpha_i k_i)(\sum \beta_i k_i) - \alpha \beta, \]

where

\[ H_i = \varepsilon \frac{\hat{x}_i^2}{S(x_{ij} - \bar{x})^2}. \]

In the case of stratified sampling

\[ \varepsilon_{ab} = M^{-1} \sum M_i \sigma^2_{e_i} m_i + M^{-1} \sum M_i \sigma^2_{x_i}, \]

\[ \varepsilon_{ab}(a, b) = -M^{-1} \sum M_i \sigma^2_{x_i} H_i. \]

For \(\varepsilon(a', b')\) we obtain when \(\beta = \cdots = \beta_N\)

\[ \varepsilon(a', b') = \varepsilon \sum k_i \hat{x}_i^2 / m_i + \varepsilon(\sum k_i \hat{x}_i^2 \{ \sum \sigma^2_{B_i} \} \} + \varepsilon \sum k_i \alpha_i, \]

\[ \varepsilon(a', b') = -\varepsilon(\sum k_i \hat{x}_i^2 \{ \sum \sigma^2_{B_i} \} \}. \]

In the general case the corresponding expressions become very involved.

5.3 UNBIASED ESTIMATION OF THE VARIANCES AND THE COVARIANCE OF THE ESTIMATED REGRESSION COEFFICIENTS

We assume that each \(m_i\) exceeds 2. When \(N = 1\) we estimate the conditional variances of \(b\) and \(a\) and their conditional covariance given \(x\) as usual by

\[ \nu(b) = \phi^2 / S(x_i - \bar{x})^2, \quad \nu(a) = \phi^2 \{ (1/m) + \bar{x}^2 / S(x_{ij} - \bar{x})^2 \}, \]

\[ \varepsilon(a, b) = -\phi \bar{x} / S(x_j - \bar{x})^2, \]

where

\[ \phi^2 = S(y_j - a - bx_j)^2 / (m - 2). \]

We now examine the case \(N > 1\) and assume \(n > 1\). Let

\[ d_{ij} = y_{ij} - \hat{a}_i - \hat{b}_i \hat{x}_{ij} \quad \hat{\sigma}_i^2 = S d_{ij}^2 / (m_i - 2), \]

then \(\hat{\sigma}_i^2 = \sigma_i^2\) and we obtain an unbiased estimate of \(\varepsilon \sum \sigma_i^2 k_i^2 F_i^2\), the variance in the stratified case, by taking

\[ \nu_{ab}(b) = \sum \hat{\sigma}_i^2 k_i^2 / S(x_{ij} - \bar{x})^2. \]

Now since

\[ \varepsilon \hat{\beta}_i^2 = \varepsilon (\beta_i + k_i^{-1} S B_i \hat{e}_{ij})^2 = \beta_i^2 + \sigma_i^2 F_i^2, \]

\[ \varepsilon \hat{\beta}_i \hat{\beta}_i' = \hat{\beta}_i \hat{\beta}_i' (i \neq i'), \]

we have

\[ \varepsilon \sum \hat{\sigma}_i^2 k_i^2 / S(x_{ij} - \bar{x})^2 = \sum p_i \sigma_i^2 k_i^2, \]

\[ \varepsilon \sum (1 - p_i) \{ \hat{\beta}_i^2 - \hat{\sigma}_i^2 k_i^2 / S(x_{ij} - \bar{x})^2 \} k_i^2 = \sum p_i (1 - p_i) \beta_i^2 k_i^2, \]

\[ \varepsilon \sum k_i \hat{\beta}_i \hat{\beta}_i' (p_i - p_i') \beta_i k_i / p_i = \sum (p_i - p_i') \beta_i \beta_i' k_i k_i'. \]

Hence, since \(\nu_{ab}\) is the sum of the right-hand sides of the last three expressions,
an unbiased estimate, $v(b)$, of $\Sigma b$ is obtained by adding the terms on the left-hand sides following the expectation signs:

$$v(b) = \sum p_i \hat{s}_i^2 / S(x_i) \sum z_i^2 + b^2 - \sum p_i \hat{s}_i^2$$

$$= \sum \sum_{i \in v} p_i \delta_i \hat{\beta}_i \hat{k}_i k_{iv} / p_{iv}.$$  \hfill (5.3.3)

When sampling with equal probabilities this gives

$$\frac{n}{N} \left\{ \sum \hat{s}_i^2 / S(x_i) \sum z_i^2 + \frac{N-n}{n} \sum \left( \hat{\beta}_i - \frac{b}{n} \right)^2 \right\}, \hfill (5.3.4)$$

which is positive. In general, however, the estimate is not always positive. In connection with an analogous situation arising in the estimation of the variance of an estimate of a mean, Sen [3] and Grundy [4] proposed an alternative method based on (5.2), which in our case leads to the unbiased estimate

$$v'(b) = \sum p_i \hat{s}_i^2 / S(x_i) \sum z_i^2$$

$$+ \frac{1}{2} \sum \sum_{i \in v} (p_i \delta_i - p_{iv})(\hat{\beta}_i - \hat{\beta}_v \hat{k}_v)^2 / p_{iv}. \hfill (5.3.5)$$

In sampling with equal probabilities $v'(b) = v(b)$. Sen [3] has shown that $p_i \delta_i$ also exceeds $p_{iv}$ if all clusters but the first are selected with equal probability. He also showed $p_i \delta_i > p_{iv}$ if $n = 2$ and if—supposing that the first cluster selected is cluster number $v$ and the second is number $v' 
eq v$—the second cluster is selected with probability $p_i^v / (1 - p_i^v)$.

Similarly we get the following unbiased estimates of $v(a, b)$ and $c_{iv}(a, b)$:

$$v_{iv}(a) = \sum \hat{s}_i^2 \{1 / m_i + \hat{x}_i / S(x_i) \sum z_i^2\}, \hfill (5.3.6)$$

$$c_{iv}(a, b) = - \sum \hat{s}_i^2 \hat{z}_i / S(x_i) \sum z_i^2; \hfill (5.3.7)$$

and more in general the following unbiased estimates of $v$ and $c(a, b)$:

$$v(a) = \sum p_i \hat{s}_i^2 \{1 / m_i + \hat{x}_i / S(x_i) \sum z_i^2\}$$

$$+ \frac{1}{2} \sum \sum_{i \in v} (p_i \delta_i - \sum \sum_{i \in v} p_i \delta_i \hat{\beta}_i \hat{k}_v k_{iv} / p_{iv}. \hfill (5.3.8)$$

$$c(a, b) = - \sum p_i \hat{s}_i^2 \hat{z}_i / S(x_i) \sum z_i^2 + ab$$

$$- \sum p_i \hat{\beta}_i \hat{k}_i - \sum \sum_{i \in v} p_i \delta_i \hat{\beta}_i \hat{k}_v k_{iv} / p_{iv}. \hfill (5.3.9)$$

The expressions for $v'(a)$ and $c'(a, b)$ are analogous.

When $\hat{\beta}_1 = \cdots = \hat{\beta}_N$ we have for unbiased estimates of $\Sigma b'$ and $c(a', b')$

$$v(b') = \sum \hat{s}_i S(B_i) \sum z_i^2, \hfill (5.3.3')$$

$$c'(a', b') = - (\sum k_i \hat{x}_i) \sum \hat{s}_i S(B_i^2). \hfill (5.3.9')$$

Under this condition no such simple estimate of $a'$ emerges; one estimate is

$$v(a') = v(a) - \sum \hat{s}_i^2 \{\hat{x}_i / S(x_i) \sum z_i^2\}$$

$$+ \sum k_i \hat{x}_i \sum \hat{s}_i S(B_i^2), \hfill (5.3.8')$$
another is obtained by replacing \( v(a) \) by \( v'(a) \) in this expression. An exception is the stratified case with
\[
v_{sr}(a) = (\sum k_i x_i)^2 \sum \sigma_i S(B_i) + \sum k_i \sigma_i / m_i. \tag{5.3.6'}
\]

6. SELECTING CLUSTERS WITH REPLACEMENT

The case in which clusters are selected with replacement will now be discussed. In this case the sequence \( v_1, \cdots, v_n \) of clusters may include some clusters more than once and \( p_r \) is less than \( p^{1}_r + \cdots + p^{n}_r = n \bar{p}_r \) (say). In general cluster number \( r \) is included \( r_r(v_1, \cdots, v_n) \) times in the random sequence \( v_1, \cdots, v_n \) of clusters and \( r \), has as possible values 0, 1, \cdots, \( n \). Let \( x'_i \) be equal to 1 if cluster number \( r \) is included at the \( i \)th draw and zero otherwise. Then \( r = \sum x'_i \) and since
\[
E x_i = p_i, \quad E x_i x'_{i'} = p_{ii'},
\]
we have
\[
E r = \varepsilon \sum (x_i), \quad E x_i x'_{i'} = n (n - 1) p_{ii'}, \tag{6.1}
\]
and for \( r \neq r' \)
\[
E r x'_{i'} = \varepsilon \sum (x_i x'_{i'}), \quad E x_i x'_{i'} = n (n - 1) p_{ii'}, \tag{6.2}
\]
since, for \( r \neq r' \), \( x_i x'_{i'} \) is necessarily zero. Here \( p_{ii'} \) is the average of all terms \( p_{ii'} \) and \( p_{ii'} \). If successive draws are independent, \( p_{ii'} = p_i p_{i'} \), and we have the so-called Poisson binomial sampling scheme in which the probabilities may change from draw to draw. Writing
\[
\omega_{ii'} = n^{-1} \sum_i (p_i - \bar{p})(p_{i'} - \bar{p}_{i'}),
\]
we have
\[
p_{ii'} = \bar{p}_r \bar{p}_{i'} - (n - 1)^{-1} \omega_{ii'}. \tag{6.4}
\]
If all \( \omega_{ii'} = 0 \), \( r \) is a binomial variable.

For later use we note the identity
\[
n \bar{p}_r + n (n - 1) \bar{p}_{ii'} - n \bar{p}_r^2 = \sum r x_r \{ n \bar{p}_r \bar{p}_{i'} - n (n - 1) \bar{p}_{ii'} \} \tag{6.5}
\]

analogous to (5.3). This I owe to Mr. N. F. Nettheim [2], and follows from the two relations
\[
\sum r x_r \bar{p}_r \bar{p}_{i'} = n \bar{p}_r \sum r x_r \bar{p}_{ii'} - n \bar{p}_r \sum r x_r - n \bar{p}_r = n \bar{p}_r - n \bar{p}_r^2 = n \bar{p}_r - n \bar{p}_r^2,
\]
\[
\sum r x_r n (n - 1) \bar{p}_{ii'} = \sum r x_r \sum r x_{i'} \bar{p}_{ii'} = \sum r x_{i'} \sum r x_{i'} \bar{p}_{ii'} = \sum r x_{i'} (n \bar{p}_{ii'} - \bar{p}_{ii'}).
\]

There are a number of possible procedures one might follow for clusters that are included in the sample more than once. In what follows we shall confine ourselves to the case where only one sample of preassigned size is taken from that cluster, but shall consider a different case in section 6.4.
We note that if \( u(r_i) \) and \( u(r_i, r_i') \) are, for any clusters \( r_i \) and \( r_i' \), defined and independent of \( R_i \) and \( R_i' \), then
\[
\varepsilon \sum u(r_i) = \sum \varepsilon_r u(r) = n \sum \bar{p}_i u(r), \tag{6.6}
\]
\[
\varepsilon \sum \sum u(r_i, r_i') = \sum \sum \varepsilon_{r,r'} u(r_i, r_i') = n(n-1) \sum \varepsilon(r, r') \bar{p}_{r,r'} u(r, r') + n(n-1) \sum \bar{p}_{r,r'} u(r, r). \tag{6.7}
\]

In case our conventions leave some doubt as to the meaning of the symbols \( \bar{p}_i, \bar{p}_{i,i'}, \bar{p}_{i,i'} \) used in what follows, let us define them explicitly. If the cluster drawn on occasion \( i \) is cluster number \( r \) and on occasion \( i' \) (\( i' > i \)) cluster number \( r' \) (different from or equal to \( r \)), then
\[
n \bar{p}_i = \sum \bar{p}_r^i, \quad n \bar{p}_{i,i'} = \sum \bar{p}_r^i, \quad n(n-1) \bar{p}_{i,i'} = \sum \bar{p}_r^i. \tag{6.8}
\]

All formulae relating to \( b' \) and \( a' \) in the previous sections also apply to sampling with replacement, with the understanding that when the method of section 6.4 is not used, \( k \) is defined by (6.1.2), and that when this method is used \( m \) is replaced by \( m' \).

### 6.1 Point Estimates of the Regression Coefficients

We compute
\[
b = \sum k \hat{\beta}_r, \quad a = \sum k \hat{\alpha}_r. \tag{6.1.1}
\]

Here summation is over the sequence of selected clusters, so that several of the \( \hat{\beta}_r, \hat{\alpha}_r \) may be identical, and (to obtain unbiasedness)
\[
k_r = M_r/(n \bar{p}_r M). \tag{6.1.2}
\]

### 6.2 The Variances of the Estimated Regression Coefficients and Their Covariance

Using the \( r_{r,i} \) we can write \( b = \sum r_{r,i} k \hat{\beta}_r \), so that
\[
\varepsilon \bar{b} = \sum \varepsilon_r \hat{\beta}_r, \quad \varepsilon b = \sum \varepsilon_{r,r'} \hat{\beta}_r \hat{\beta}_{r'} = \sum \varepsilon_r \hat{\beta}_r + (n-1) \bar{p}_r \varepsilon \hat{\beta}_r. \tag{6.2.1}
\]

For the variance
\[
\nu b = \sum r_{r,i} \hat{\beta}_r k_r = \sum (r_{r,i}^2 \hat{\beta}_r^2 k_r^2) + \sum \varepsilon_{r,r'} \hat{\beta}_r \hat{\beta}_{r'} k_r k_{r'}
\]
\[
= n \sum \{ \bar{p}_r + (n-1) \bar{p}_{r'} - n \bar{p}_{r} \bar{p}_{r'} \} \hat{\beta}_r \hat{\beta}_{r'} k_r k_{r'}
\]
\[
+ n \sum \varepsilon_{r,r'} k_r k_{r'} \{ (n-1) \bar{p}_{r'} - n \bar{p}_r \bar{p}_{r'} \} \hat{\beta}_r \hat{\beta}_{r'} k_r k_{r'}
\]
\[
= n \sum \bar{p}_r \hat{\beta}_r^2 k_r^2 + (n-1) \sum \varepsilon_{r,r'} \bar{p}_{r'} \hat{\beta}_{r'} k_r k_{r'} - \beta^2. \tag{6.2.2}
\]

and \( \nu b \) is the sum of these two expressions.

Similarly
\[
\nu a = n \sum \{ \bar{p}_r + (n-1) \bar{p}_{r'} \} \hat{\beta}_r \hat{\beta}_{r'} k_r k_{r'} + \sum \bar{p}_r \hat{\alpha}_r^2 k_r^2
\]
\[
+ n(n-1) \sum r_{r,i} \hat{\beta}_r \hat{\alpha}_r k_r k_{r'} - \alpha^2. \tag{6.2.3}
\]
\[ c(a, b) = - \sum \left\{ \bar{p}_r + (n - 1) \bar{p}_r \right\} \sigma^2_{\bar{p}_r} + n \sum \bar{p}_r \sigma_{\bar{p}_r} \hat{\sigma}_{r}^2 \]

\[ + n(n - 1) \sum \sum_{r', v} \bar{p}_{r'v} \sigma_{\bar{p}_{r'v}} \hat{\sigma}_{r'}^2 \]

6.3 Unbiased Estimates of the Variances and Covariance of the Estimated Regression Coefficients

Applying (6.7) with
\[ v(i, i') = \left\{ (n - 1) \beta_{iv} - n \bar{p}_i \bar{p}_i \right\} \sigma_{\bar{p}_i}^2 / \left\{ (n - 1) \bar{p}_i \right\}, \]

\[ w(i, i') = \epsilon_{iv} \beta_{iv}, \quad w(i, i') = \hat{\beta}_i + \sigma^2_{\bar{p}_i}, \]

gives that

\[ \sum \sum_{i,i'} \left( \beta_{ii'} \right) \bar{p}_{i'v} \]

\[ = n \sum \sum_{i,i'} \left( (n - 1) \bar{p}_{i'v} - n \bar{p}_i \bar{p}_v \right) \beta_{ii'} \sigma_{\bar{p}_i} \hat{\sigma}_{i'}^2 \]

\[ + n \sum \left( (n - 1) \bar{p}_{i'v} - n \bar{p}_v \right) \left( \beta_{ii'}^2 + \sigma^2_{\bar{p}_i} \hat{\sigma}_{i'}^2 \right). \]  

On the other hand, use of (6.6) gives

\[ \sum \left( (n - 1) \bar{p}_{ii} - n \bar{p}_i \bar{p}_i \right) \left( \beta_{ii'}^2 + \sigma^2_{\bar{p}_i} \hat{\sigma}_{i'}^2 \right) \]

so that an unbiased estimate of (6.2.2) is

\[ \sum \left( \bar{p}_i + (n - 1) \bar{p}_i \right) \left( \beta_{ii'} \right) \sigma^2_{\bar{p}_i} / \left\{ (S(x_{i1} - \bar{x}_i)^2 \right\} \]

\[ + \sum \sum_{i,i'} v(i, i') \beta_{ii'} \beta_{i'v} - \sum \left( (n - 1) \bar{p}_{ii} - n \bar{p}_i \right) \left( \beta_{ii'} \right) \sigma^2_{\bar{p}_i} \hat{\sigma}_{i'}^2. \]  

An unbiased estimate of (6.2.1) is

\[ \sum \left( \bar{p}_i + (n - 1) \bar{p}_i \right) \sigma^2_{\bar{p}_i} \]

so that we get

\[ v(b) = \sum \left( n \bar{p}_i \sigma^2_{\bar{p}_i} / \left\{ (S(x_{i1} - \bar{x}_i)^2 \right\} \]

\[ - \frac{n}{n - 1} \sum \sum_{i,i'} \bar{p}_i \bar{p}_v \beta_{ii'} \sigma_{\bar{p}_i} \hat{\sigma}_{i'}^2 \bar{p}_{i'v}. \]  

Similarly,

\[ v(a) = \sum \left( \bar{p}_i \sigma^2_{\bar{p}_i} \left\{ (1/m_i) + \bar{x}_i / (S(x_{i1} - \bar{x}_i)^2 \right\} \]

\[ + a^2 - \frac{n}{n - 1} \sum \sum_{i,i'} \bar{p}_i \bar{p}_v \sigma_{\bar{p}_i} \hat{\sigma}_{i'}^2 \bar{p}_{i'v}. \]  

\[ c(a, b) = - n \sum \left( n \bar{p}_i \sigma^2_{\bar{p}_i} \left\{ (S(x_{i1} - \bar{x}_i)^2 \right\} + \sigma^2_{\bar{p}_i} \right) \]

\[ - \frac{n}{n - 1} \sum \sum_{i,i'} \bar{p}_i \bar{p}_v \sigma_{\bar{p}_i} \hat{\sigma}_{i'}^2 \bar{p}_{i'v}. \]  

Substitution of (6.5) in (6.2.2) gives for \( \forall_{\bar{p}_{r1}, \ldots, \bar{p}_b} \)

\[ n \sum \sum_{r,r'} \left\{ n \bar{p}_r \bar{p}_{r'} - (n - 1) \bar{p}_{r'r} \right\} \left\{ \beta_{r'r} \right\} \sigma^2_{\bar{p}_r \bar{p}_{r'}} \]

\[ = \frac{1}{n} \sum \sum_{r,r'} \left\{ n \bar{p}_r \bar{p}_{r'} - (n - 1) \bar{p}_{r'r} \right\} \left( \beta_{r'r} - \hat{\beta}_{r'r} \right)^2. \]
and leads to the estimate
\[ v'(b) = n \sum \hat{p}_i \hat{d}_i^2 \frac{1}{S(x_{ij} - \bar{x}_i)^2} \]
\[ + \frac{1}{2} \sum \sum_{i < j} \{ n \hat{p}_i \hat{d}_i - (n - 1) \hat{p}_i \hat{d}_j \} (\hat{\beta}_i \hat{k}_i - \hat{\beta}_j \hat{k}_j)^2 / \{ (n - 1) \hat{p}_i \hat{d}_j \}. \] (6.3.5)
For by (6.7) the mean of the second term in (6.3.5) is
\[ \frac{1}{2} n \sum \sum_{i < j} \{ n \hat{p}_i \hat{d}_i - (n - 1) \hat{p}_i \hat{d}_j \} (\hat{\beta}_i \hat{k}_i - \hat{\beta}_j \hat{k}_j)^2 \]
\[ + \frac{1}{2} n \sum \sum_{i < j} \{ n \hat{p}_i \hat{d}_i - (n - 1) \hat{p}_i \hat{d}_j \} (\sigma_i^2 \hat{\beta}_i^2 + \sigma_j^2 \hat{\beta}_j^2) \]
and by (6.5) the second term in this sum equals
\[ \sum \{ n \hat{p}_i + n(n - 1) \hat{p}_i - n \hat{p}_i^2 \} \sigma_i^2 \hat{\beta}_i^2. \]
But this is smaller than (6.2.1) by
\[ \sum n \hat{p}_i \hat{d}_i^2 \frac{1}{S(x_{ij} - \bar{x}_i)^2}. \]
which may be estimated by
\[ \sum n \hat{p}_i \hat{d}_i^2 \frac{1}{S(x_{ij} - \bar{x}_i)^2}. \]
The expressions for \( v'(a) \) and \( c'(a, b) \) are similar.

6.4 ESTIMATES FOR AN ALTERNATIVE WAY OF SUBSAMPLING

Instead of taking a subsample of preassigned size \( m \), from cluster number \( \nu \) if this cluster is included at least once, we may take a sample of size
\[ m' = \min (r, m, M). \]
In this section we shall consider estimation problems for this case. For typographical simplicity we shall make a minimum of notational changes; this cannot lead to confusion as the present case is discussed in the present section only.

If \( r'_\nu = r'_\nu(v_1, \ldots, v_n) \) is 1 if cluster \( \nu \) is included in the sequence \( v_1, \ldots, v_n \), and 0 otherwise, we obtain the following estimates for \( \hat{\beta} \) and \( \hat{\alpha} \):
\[ b = \sum k_i r'_\nu \hat{\beta}_i, \quad a = \sum k_i r'_\nu \hat{\alpha}_i, \] (6.4.1)
which are unbiased only when
\[ k_i = M_i / (p_i M). \] (6.4.2)
Writing
\[ F'_\nu (v_1, \ldots, v_n) = \epsilon_{v_1 \ldots v_n} 1 / S(x_{ij} - \bar{x}_i)^2, \]
we have
\[ \epsilon \sum \epsilon_{v_1 \ldots v_n} b = \sum k_i \hat{d}_i^2 \{ r'_\nu (v_1, \ldots, v_n) F'_\nu (v_1, \ldots, v_n) \}. \] (6.4.3)
From the fact that \( r'_\nu (v_1, \ldots, v_n) \) is, for any \( \nu \), a binomial variate with mean \( p_\nu \), and \( r'_\nu (v_1, \ldots, v_n) r'_\nu (v_1, \ldots, v_n) \), for any pair of different \( \nu \) and \( \nu' \), a binomial variate with mean \( p_{\nu \nu'} \), we obtain
\[ \sum_{b} = \sum_{p} \beta_{b} \beta_{b'} - \sum_{p} \sum_{p'} \beta_{p} \beta_{p'} \]

and \( \bar{b} \) is the sum of (6.4.3) and (6.4.4).

As

\[ \varepsilon_{b} = \sum_{p} \beta_{b} \beta_{b'} (v_{1}, \ldots, v_{n}) \]

\[ = \sum_{p} \beta_{b} \beta_{b'} (v_{1}, \ldots, v_{n}) \frac{r'_{b} (v_{1}, \ldots, v_{n})}{r'_{b'} (v_{1}, \ldots, v_{n})} \]

has mean

\[ \sum_{p} \beta_{b} \beta_{b'} (v_{1}, \ldots, v_{n}) \frac{r'_{b} (v_{1}, \ldots, v_{n})}{r'_{b'} (v_{1}, \ldots, v_{n})} \]

and

\[ \varepsilon_{b} = \sum_{p} \beta_{b} \beta_{b'} (v_{1}, \ldots, v_{n}) \frac{r'_{b} (v_{1}, \ldots, v_{n})}{r'_{b'} (v_{1}, \ldots, v_{n})} \]

has mean

\[ \sum_{p} \beta_{b} \beta_{b'} (v_{1}, \ldots, v_{n}) \frac{r'_{b} (v_{1}, \ldots, v_{n})}{r'_{b'} (v_{1}, \ldots, v_{n})} \]

an unbiased estimate of \( \bar{b} \) is

\[ \bar{v}(b) = \sum_{p} \beta_{b} \beta_{b'} (v_{1}, \ldots, v_{n}) / \sum_{p} r'_{b} (v_{1}, \ldots, v_{n}) \]

\[ + \sum_{p} \beta_{b} \beta_{b'} (v_{1}, \ldots, v_{n}) \]

\[ - \sum_{p} \sum_{p'} \beta_{p} \beta_{p'} (v_{1}, \ldots, v_{n}) r'_{b} (v_{1}, \ldots, v_{n}) / \sum_{p} \sum_{p'} \beta_{p} \beta_{p'} (v_{1}, \ldots, v_{n}) \]

The results for \( \bar{a}, \bar{c}(a, b), \bar{v}(a) \) and \( \bar{a} \) which are certain to be positive in a number of important cases. However, if the abovementioned difficulties are serious, we can still compute the estimates of the previous section by using only a random subset of \( m \), observations from cluster \( a \).

7. RESULTS FOR THE SECOND MODEL

In this section we briefly outline some of the results based on the second model. It is easy to see that the means of \( a, b, a', \) and \( b' \) are the same in both models. The variances of \( a, b, \) and \( b' \) and the covariances of \( a, b, \) and \( a' \) and \( b' \) are the same as given in sections 5.2 and 6.2, provided one replaces \( \sigma_{a}^{2} \) by \( \sigma_{a'}^{2} - \tau_{a} \). Moreover, the estimates of these quantities given in sections 5.3 and 6.3 remain valid, since the mean of \( \sigma_{a}^{2} \) is \( \sigma_{a'}^{2} - \tau_{a} \) and since the conditional mean of \( \sigma_{b}^{2} \) is unchanged provided \( \sigma_{b}^{2} \) is replaced by \( \sigma_{b'}^{2} - \tau_{b} \). To obtain the variances of \( a \) and \( a' \) one has to make not only the replacement of \( \sigma_{a}^{2} \) by \( \sigma_{b}^{2} - \tau_{a} \), but also has to add the term \( \sum_{i} \beta_{i} \beta_{i}^{2} \). It then turns out that the estimates of the variance of \( a \) (or \( a' \) discussed in sections 5.3 and 6.3 have a mean value too small by \( \sum_{i} \beta_{i} \beta_{i}^{2} \tau_{a} \) and \( \sum_{i} \beta_{i} \beta_{i}^{2} \tau_{a} \), respectively. We do not know how to estimate the \( \tau_{a} \) and \( \bar{a}^{2} \) can be obtained as follows:
\[ e_i = \tau_i + \alpha_i + \beta_i(x_{ij} + x_{i}) + \xi_i \]
\[ e_i = \alpha_i \beta_i / S(x_{ij} - \bar{x}) ^ 2 \]
\[ e_i = \beta_i / S(x_{ij} - \bar{x}) ^ 2 \]
and
\[ SS_{ij} = \gamma_i m_i (m_i - 1) - S(y - \bar{y}) ^ 2 \]

so
\[ \gamma_i = \left\{ m_i (m_i - 1) \right\} ^ {-1} S(y - \bar{y}) ^ 2 - 2 \left\{ \beta_i / S(x_{ij} - \bar{x}) ^ 2 \right\} \bar{x}_i 
+ (m_i - 1) ^ {-1} \left\{ \beta_i S(x_{ij} - \bar{x}) ^ 2 - \beta_i \right\} (m_i - 1) / S(x_{ij} - \bar{x}) ^ 2 - \bar{x}_i \]

has a mean (given \( i \)) of
\[ \tau_i + \alpha_i \]

Perhaps it is reasonable to assume that the intraclass correlation \( \tau_i / \sigma_i ^ 2 \) is constant for all or for groups of the clusters. Consider, for example, the case where this correlation is assumed to be constant, say \( \rho \), and let us assume that for two of the clusters, say 1 and 2, the \( \alpha \)'s are the same. Then the difference, \( t_i = t_2 \), between the values of (7.1) for clusters 1 and 2 estimates \( \tau_1 - \tau_2 = \rho (\alpha_1 - \alpha_2) \).

Now \( \delta^2 / \sigma^2 \) estimates \( (\delta^2 / \sigma^2) (1 - \rho) \), so \( \rho \) may be estimated by
\[ \frac{t_1 - t_2}{t_1 - t_2 + \delta_1^2 - \delta_2^2} \]

and \( \tau_i \) by
\[ \tau_i = \frac{\delta_i^2 (t_1 - t_2)}{\delta_1^2 - \delta_2^2} \]

8. DISCUSSION

One is often interested in estimating an average relationship between certain variables in a given population. This paper was written to promote wider use of surveys for that purpose. This use can be two-fold. Surveys taken in the past to estimate averages or totals may contain considerable information on such relationships. From such surveys, the various magnitudes entering into the formulas for the variances and covariances of estimates could be established, and these may help in designing new surveys which yield good estimates of regression relations or of such relations coupled with averages and totals.

Analyses of past surveys could also throw light on the adequacy of the proposed model.

As mentioned in section 2, generalization to multiple regression is straightforward. A very simple special case of much interest is one in which, while still requiring an average line of relationship, we allow this line to be at differ-
ent levels for different strata (or aggregates of strata). Let there be \( D \) such strata and let \( q_d \) for an individual be 1 or 0 according to whether the individual is or is not in stratum \( d \); the desired regression is \( y = \alpha_0 q_1 + \cdots + \alpha_D q_D + \beta x + z \). If we denote the estimated regression in stratum \( d \) by \( \hat{a}_d + \hat{b}_d x \), the \( \hat{a}_d \) estimate the \( a_d \) and the estimate of \( \beta \) is obtained by combining the \( \hat{b}_d \) as before.

REFERENCES


APPENDIX: GLOSSARY OF SYMBOLS

(Numbers in parentheses indicate equations. Symbols used in one section only are excluded.)

\( \alpha, \beta \) regression coefficients in the overall regression equations.
\( \alpha, \beta \) unbiased estimates of \( \alpha \) and \( \beta \): (5.1.7), (5.1.2).
\( \alpha', \beta' \) ratio estimates of \( \alpha \) and \( \beta \): (5.1.6'), (5.1.3').
\( \alpha_i, \beta_i \) regression coefficients in class \( i \) (first model) or for each individual in class \( i \) (second model).
\( \hat{\alpha}_i, \hat{\beta}_i \) least squares estimates of \( \alpha_i \) and \( \beta_i \): (5.1.5), (5.1.1).
\( A_{ij}, B_{ij} \) (5.1.6), (5.1.3).
\( A'_{ij}, B'_{ij} \) (5.1.6'), (5.1.3').
\( c_i \) arbitrary factor in definitions of \( A'_{ij} \) and \( B'_{ij} \).
\( \epsilon, c(\ldots) \) covariance and estimate thereof; subscript str indicates stratified sample.
\( \epsilon \) expected value, subscripts indicate conditional expected values, as defined in section 6.2.
\( F_i \) (5.2.1).
\( G_i \) (5.2.8).
\( H_i \) (5.2.10).
\( i \) runs from 1 to \( n \).
\( j \) runs from 1 to \( m_i \).
\( k_i \) (5.1.4) for sampling without replacement; (6.1.2) for sampling with replacement (except in section 6.4).
\( M, M \) number of individuals in entire population and in class \( v \), respectively.
\( m_i \) number of individuals in sample for class \( i \).
\( N, n \) number of classes in population and in sample, respectively.
\( \nu \) runs from 1 to \( N \).
\( p \) with subscripts and superscripts defined in section 4.
\( \bar{\nu}, \bar{\nu}^*, \bar{\nu}^{**} \) defined for sampling with replacement as the average of the \( \bar{\nu}^i, \bar{\nu}^{*i}, \) and \( \bar{\nu}^{**i} \).
\(\tilde{p}_i, \tilde{p}_i, \tilde{p}_i, \) (6.8).

\(R_i\) sample from class \(i\).

\(\tau_r(\nu_1, \cdots, \nu_n)\) number of times cluster number \(r\) is included in the random sequence \(\nu_1, \cdots, \nu_n\) of clusters.

\(\Sigma\) summation over \(\nu\) or \(i\).

\(S\) summation over \(j\).

\(\sigma_i^2\) conditional variance of residual in class \(i\) (first model) or of each individual in class \(i\) (second model).

\(\delta_i^2\) (5.3.1); unbiased estimate of \(\sigma_i^2\) (first model) or \(\sigma_i^2 - \tau_i\) (second model).

\(\tau_i\) conditional covariance of residual for any two individuals in class \(i\) (second model).

\(\mathbb{V}, v(\ )\) variance and estimate thereof; subscript str indicates stratified sample; other subscripts indicate conditional variances, as defined in section 6.2.

\(x, y, z\) regressor, regressand, and residual; used with and without subscripts.