CHAPTER 20

Complex Survey Design: Regression Analysis

20.1 INTRODUCTION

In regression analysis we describe the relationship between a response (dependent) variable and a number of explanatory (independent) variables. We also predict the future value of the dependent variable using the established relationship. The relationships are explained through variances, simple and multiple correlations, regression coefficients, fitting regression of one variable on others, and so on. For example, in BAIS II survey one may be interested in finding the relationship between HIV infection rates and age, sex, economic and social conditions, etc. The classical method of regression analysis is based on the assumption that the data are collected through simple random sampling with replacement method and software packages such as SAS and SPSS are based on this assumption. But in reality, large-scale surveys are generally based on complex survey designs involving stratification, clustering, and unequal probability of selection of samples. Therefore the assumption of independency of observations is rarely valid; hence the output of the standard software packages are erroneous or misleading when the underlying sampling design is ignored. In general, assumption of independency of observations underestimates the variance and hence underestimates confidence intervals also. Kish and Frankel (1974) recommended that "Standard errors should be computed in accordance with the complexity of the sampling designs; neglect of that complexity is a common source of serious mistakes." In this chapter we will consider regression analysis from complex survey designs under design-based, model-based, and model/design-based approaches. In regression analysis, we first indentify the appropriate model and parameters for inference and then we need to consider the type of inference that is required viz. point or interval estimation, or testing of hypotheses (see Nathan, 1988).

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00020-0

20.2 DESIGN-BASED APPROACH

In a design-based approach, we consider a finite population $U = (U_1,\ldots, U_i,\ldots, U_N)$ of N units. Associated with the unit U_i, we have y_i, a study variable, and $\mathbf{x}_i = (x_{0i}, x_{1i},\ldots, x_{ji},\ldots, x_{pi})$, a vector of p auxiliary (independent) variables $x_1,\ldots, x_p$ with $x_{0i} = 1$. Here our parameters of interest (see Nathan, 1988) B_0, $B_1,\ldots, B_j,\ldots, B_p$ are such that

$$\sum_{i=1}^{N}\left(y_i - B_0 - B_1x_{1i} - \cdots - B_jx_{ji} - \cdots B_px_{pi}\right)^2 \tag{20.2.1}$$

attains a minimum, i.e., parameter of interest $\mathbf{B} = (B_0, B_1, \ldots, B_p)'$ is the ordinary least square (OLS) estimate if the whole population is observed. Here the regression coefficient $\mathbf{B}$ is a descriptive parameter because its value can be determined without any error if the entire population is surveyed and if no response and measurement error is present. Similarly, the finite population mean, variance, and correlation coefficient are also descriptive parameters. In contrast, the model parameters such as mean and variance of a normal distribution can never be calculated exactly.

Let $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \cdot \\ \mathbf{x}_i \\ \cdot \\ \mathbf{x}_N \end{pmatrix}$ be of full rank and $\mathbf{y} = \begin{pmatrix} y_1 \\ \cdot \\ y_i \\ \cdot \\ y_N \end{pmatrix}$, then the solution of Eq. (20.2.1) is

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{20.2.2}$$

Denoting $T_{yj} = \sum_{i\in U} y_i x_{ji}$ and $T_{jk} = \sum_{i\in U} x_{ji}x_{ki}$ for j, $k = 0,\ldots, p$, Eq. (20.2.2) can be written as

$$\mathbf{B} = \mathbf{T}_{xx}^{-1}\mathbf{T}_{xy} \tag{20.2.3}$$

where $\mathbf{T}_{xx} = (\mathbf{X}'\mathbf{X})$ is a $(p+1)\times(p+1)$ matrix with the jth and kth element as T_{jk} and $\mathbf{T}_{xy} = \mathbf{X}'\mathbf{y}$ is a $(p+1)\times 1$ column matrix with jth element T_{yj}.

To estimate $\mathbf{B}$, we select a sample s of size n from the population U with probability $p(s)$ by using a suitable sampling design p. Let π_i (>0) and π_{ij} (>0) denote the inclusion probabilities for the ith, and ith and jth units. Let the quantities T_{yj} and T_{jk} be estimated unbiasedly using the estimators

$$\widehat{T}_{yj} = \sum_{i \in s} b_{si} y_i x_{ji} \text{ and } \widehat{T}_{jk} = \sum_{i \in s} b_{si} x_{ji} x_{ki} \tag{20.2.4}$$

where b_{si}'s are suitably chosen constants satisfying the unbiasedness condition $\sum_{s \supset i} b_{si} p(s) = 1$ for $i = 1, \ldots, N$.

Without loss of generality, let us suppose that the sample s consists of first n units with labels $1, 2, \ldots, n$. Then we can denote

$$\widehat{T}_{wyj} = (x_{j1}, x_{j2}, \ldots, x_{jn}) \mathbf{W}_s \mathbf{y}_s \text{ for } j = 0, 1, \ldots, p$$

where $\mathbf{y}_s = (y_1, \ldots, y_n)'$ and $\mathbf{W}_s = \mathbf{diag}(b_{s1}, \ldots, b_{sn})$.

In case $\mathbf{X}_s = \begin{pmatrix} \mathbf{x}_1 \\ \cdot \\ \cdot \\ \mathbf{x}_n \end{pmatrix}$ is a full rank, we can set an estimator of $\mathbf{B}$ as

$$\widehat{\mathbf{B}}_w = \widehat{\mathbf{T}}_{wxx}^{-1} \widehat{\mathbf{T}}_{wxy} \tag{20.2.5}$$

where $\widehat{\mathbf{T}}_{wxx} = \mathbf{X}_s' \mathbf{W}_s \mathbf{X}_s$ is a $(p+1) \times (p+1)$ matrix with jth and kth elements as $\widehat{T}_{wjk}$ and $\widehat{\mathbf{T}}_{wxy} = \mathbf{X}_s' \mathbf{W}_s \mathbf{y}_s$ is a $(p+1) \times 1$ column matrix with the jth element $\widehat{T}_{wyj}$.

In particular, if we take $b_{si} = 1/\pi_i$, then Eq. (20.2.5) reduces to

$$\widehat{\mathbf{B}}_\pi = \widehat{\mathbf{T}}_\pi^{-1} \widehat{\mathbf{T}}_{\pi y} \tag{20.2.6}$$

where $\widehat{\mathbf{T}}_\pi = \mathbf{X}_s' \mathbf{W}_\pi \mathbf{X}_s$, $\mathbf{W}_\pi = \mathbf{diag}(1/\pi_1, \ldots, 1/\pi_n)$ and $\mathbf{T}_{\pi y}$ is a $(p+1) \times 1$ column matrix with jth element $\widehat{T}_{\pi yj} = \sum_{i \in s} y_i x_{ji} / \pi_i$.

Obviously, the estimators $\widehat{\mathbf{B}}_w$ and $\widehat{\mathbf{B}}_\pi$ are not a design unbiased in general but they are design consistent for $\mathbf{B}$.

Example 20.2.1

In case of a single auxiliary variable ($p = 1$), we find

$$B_0 = \overline{Y} - B_1 \overline{X}_1 = T_{y0}/T_{00} - B_1 T_{11}/T_{00} \text{ and}$$

$$B_1 = \frac{\sum_{i \in U} y_i x_{1i} - \left(\sum_{i \in U} x_{1i}\right)\left(\sum_{i \in U} y_i\right) \Big/ N}{\sum_{i \in U} x_{1i}^2 - \left(\sum_{i \in U} x_{1i}\right)^2 \Big/ N} = \frac{T_{y1} - T_{01} T_{y0}/T_{00}}{T_{11} - T_{01}^2/T_{00}} \tag{20.2.7}$$

where $T_{y0} = \sum_{i \in U} y_i$, $T_{y1} = \sum_{i \in U} y_i x_{1i}$, $T_{00} = N$, $T_{01} = \sum_{i \in U} x_{1i}$, and $T_{11} = \sum_{i \in U} x_{1i}^2$.

Now, writing $\widehat{T}_{w00} = \sum_{i \in s} b_{si}$, $\widehat{T}_{w01} = \sum_{i \in s} b_{si} x_{1i}$, $\widehat{T}_{w11} = \sum_{i \in s} b_{si} x_{1i}^2$, $\widehat{T}_{wy0} = \sum_{i \in s} b_{si} y_i$, and $\widehat{T}_{wy1} = \sum_{i \in s} b_{si} x_{1i} y_i$, the estimators of B_0 and B_1 are obtained as

$$\widehat{B}_{w0} = \widehat{T}_{wy0} \Big/ \widehat{T}_{w00} - \widehat{B}_{w1} \widehat{T}_{w01} \Big/ \widehat{T}_{w00} = \frac{\sum_{i \in s} b_{si} y_i}{\sum_{i \in s} b_{si}} - \widehat{B}_{w1} \frac{\sum_{i \in s} b_{si} x_{1i}}{\sum_{i \in s} b_{si}} \tag{20.2.8}$$

and

$$\begin{aligned}\widehat{B}_{w1} &= \frac{\widehat{T}_{wy1} - \widehat{T}_{w01} \widehat{T}_{wy0} \Big/ \widehat{T}_{w00}}{\widehat{T}_{w11} - \widehat{T}_{w01}^2 \Big/ \widehat{T}_{w00}} \\ &= \frac{\sum_{i \in s} b_{si} y_i x_{1i} - \left(\sum_{i \in s} b_{si} x_{1i}\right)\left(\sum_{i \in s} b_{si} y_i\right) \Big/ \left(\sum_{i \in s} b_{si}\right)}{\sum_{i \in s} b_{si} x_{1i}^2 - \left(\sum_{i \in s} b_{si} x_{1i}\right)^2 \Big/ \left(\sum_{i \in s} b_{si}\right)}\end{aligned} \tag{20.2.9}$$

In case, $b_{si} = 1/\pi_i$, the estimators (20.2.8) and (20.2.9) reduce to

$$\begin{aligned}&\widehat{B}_{\pi 0} = \left(\sum_{i \in s} \frac{y_i}{\pi_i}\right) \Big/ \left(\sum_{i \in s} \frac{1}{\pi_i}\right) - \widehat{B}_{\pi 1} \left(\sum_{i \in s} \frac{x_{1i}}{\pi_i}\right) \Big/ \left(\sum_{i \in s} \frac{1}{\pi_i}\right) \text{ and} \\ &\widehat{B}_{\pi 1} = \frac{\sum_{i \in s} \frac{y_i x_{1i}}{\pi_i} - \left(\sum_{i \in s} \frac{y_i}{\pi_i}\right)\left(\sum_{i \in s} \frac{x_{1i}}{\pi_i}\right) \Big/ \sum_{i \in s} \frac{1}{\pi_i}}{\sum_{i \in s} \frac{x_{1i}^2}{\pi_i} - \left(\sum_{i \in s} \frac{x_{1i}}{\pi_i}\right)^2 \Big/ \sum_{i \in s} \frac{1}{\pi_i}}\end{aligned} \tag{20.2.10}$$

For SRSWOR sampling $\pi_i = n/N$ and we get

$$\begin{aligned}&\widehat{B}_0 = \widehat{B}_{\pi 0} = \overline{y}_s - \widehat{B}_{\pi 1}\, \overline{x}_s \text{ and } \widehat{B}_1 = \widehat{B}_{\pi 1} \\ &= \frac{\sum_{i \in s} y_i x_{1i} - \left(\sum_{i \in s} y_i\right)\left(\sum_{i \in s} x_{1i}\right) \Big/ n}{\sum_{i \in s} x_{1i}^2 - \left(\sum_{i \in s} x_{1i}\right)^2 \Big/ n}\end{aligned} \tag{20.2.11}$$

20.2.1 Estimation of Variance

The variance of $\widehat{\mathbf{B}}_w$ can be estimated by using standard variance estimation techniques for complex survey designs such as the random group, balanced repeated replication, jackknife, and bootstrap methods. Shah et al. (1977) and Binder (1983) derived an approximate expression of variance and its estimator by using Taylor series expansion as follows.

Let us write

$$\begin{aligned}\mathbf{W}(\mathbf{B}) &= (W_1(\mathbf{B}), \ldots, W_{p+1}(\mathbf{B}))' \\ &= \sum_{i=1}^{N}(y_i - \mathbf{x}_i\mathbf{B})\mathbf{x}_i' = \mathbf{0}\end{aligned} \tag{20.2.12}$$

Then an unbiased estimator of $\mathbf{W}(\mathbf{B})$ when $\mathbf{B}$ is known is

$$\widehat{\mathbf{W}}(\mathbf{B}) = \sum_{i\in s} b_{si}(y_i - \mathbf{x}_i\mathbf{B})\mathbf{x}_i'. \tag{20.2.13}$$

Now, writing $\widehat{e}_i = y_i - \mathbf{x}_i\widehat{\mathbf{B}}_w$, we find an approximate estimator of $\widehat{\mathbf{W}}(\mathbf{B})$ as

$$\widehat{\mathbf{W}}\left(\widehat{\mathbf{B}}_w\right) = \sum_{i\in s} b_{si}\widehat{e}_i\mathbf{x}_i' \tag{20.2.14}$$

Taylor series expansion of $\widehat{\mathbf{W}}\left(\widehat{\mathbf{B}}_w\right)$ around the point $\widehat{\mathbf{B}}_w = \mathbf{B}$ yields

$$\widehat{\mathbf{W}}\left(\widehat{\mathbf{B}}_w\right) \simeq \widehat{\mathbf{W}}(\mathbf{B}) + \frac{\partial\widehat{\mathbf{W}}(\mathbf{B})}{\partial\mathbf{B}}\left(\widehat{\mathbf{B}}_w - \mathbf{B}\right) \tag{20.2.15}$$

where $\dfrac{\partial\hat{\mathbf{W}}\left(\hat{\mathbf{B}}_w\right)}{\partial\mathbf{B}}$ is a $(p+1)\times(p+1)$ matrix whose i, jth element is the partial derivatives $\dfrac{\partial\hat{W}_i\left(\hat{\mathbf{B}}_w\right)}{\partial B_j}$.

Now, writing $\widehat{\mathbf{W}}\left(\widehat{\mathbf{B}}_w\right) = 0$, we obtain

$$\widehat{\mathbf{W}}(\mathbf{B}) \cong -\frac{\partial\widehat{\mathbf{W}}(\mathbf{B})}{\partial\mathbf{B}}\left(\widehat{\mathbf{B}}_w - \mathbf{B}\right)$$

Taking variances of both sides, and taking limit (following Binder, 1983), we get

$$V\left[\widehat{\mathbf{W}}(\mathbf{B})\right] \cong \left(\frac{\partial\mathbf{W}(\mathbf{B})}{\partial\mathbf{B}}\right)V\left(\widehat{\mathbf{B}}_w\right)\left(\frac{\partial\mathbf{W}(\mathbf{B})}{\partial\mathbf{B}}\right)' \tag{20.2.16}$$

Now, assuming $\dfrac{\partial \mathbf{W}(\mathbf{B})}{\partial \mathbf{B}}$ is of full rank, we find

$$V\left(\widehat{\mathbf{B}}_w\right) \simeq \left(\frac{\partial \mathbf{W}(\mathbf{B})}{\partial \mathbf{B}}\right)^{-1} V\left(\widehat{\mathbf{W}}(\mathbf{B})\right)\left(\frac{\partial \mathbf{W}(\mathbf{B})}{\partial \mathbf{B}}\right)^{\prime -1} \tag{20.2.17}$$

Now, noting $\dfrac{\partial \mathbf{W}(\mathbf{B})}{\partial \mathbf{B}} = -\mathbf{X}'\mathbf{X}$ and $\widehat{\mathbf{W}}(\mathbf{B}) = \sum_{i \in s} b_{si} e_i \mathbf{x}_i'$ with $e_i = y_i - \mathbf{x}_i \mathbf{B}_w$, we get

$$V\left(\widehat{\mathbf{B}}_w\right) \simeq (\mathbf{X}'\mathbf{X})^{-1} V\left(\sum_{i \in s} b_{si} e_i \mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{20.2.18}$$

Finally, we get an approximate expression of $V\left(\widehat{\mathbf{B}}_w\right)$ as

$$\widehat{V}\left(\widehat{\mathbf{B}}_w\right) = \widehat{\mathbf{T}}_{wxx}^{-1}\, \widehat{\Sigma}_{ex}\, \widehat{\mathbf{T}}_{wxx}^{-1} \tag{20.2.19}$$

where $\widehat{\mathbf{T}}_{wxx}$ given in Eq. (20.2.5) is an unbiased estimator of $\mathbf{X}'\mathbf{X}$ and $\widehat{\boldsymbol{\Sigma}}_{ex}$ is an unbiased estimator of $V\left(\sum_{i \in s} b_{si} e_i \mathbf{x}_i'\right)$.

In particular, if $b_{si} = 1/\pi_i$, the variance of $\widehat{\mathbf{B}}_\pi$ becomes

$$V\left(\widehat{\mathbf{B}}_\pi\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{Q}(\mathbf{X}'\mathbf{X})$$

where $\mathbf{Q}$ is a $(p+1) \times (p+1)$ with j, kth element

$$Q_{jk} = \frac{1}{2}\sum_{i \neq}\sum_{t \in U}(\pi_i \pi_t - \pi_{it})\left(\frac{e_i x_{ji}}{\pi_i} - \frac{e_t x_{jt}}{\pi_t}\right)\left(\frac{e_i x_{ki}}{\pi_i} - \frac{e_t x_{kt}}{\pi_t}\right).$$

A consistent estimator of $V\left(\widehat{\mathbf{B}}_\pi\right)$ is

$$\widehat{V}\left(\widehat{\mathbf{B}}_\pi\right) = (\mathbf{X}'\mathbf{X})^{-1}\widehat{\mathbf{Q}}(\mathbf{X}'\mathbf{X})$$

where j, k element of $\widehat{\mathbf{Q}}$ is

$$\widehat{Q}_{jk} = \frac{1}{2}\sum_{i \neq}\sum_{t \in s}\frac{(\pi_i \pi_t - \pi_{it})}{\pi_{it}}\left(\frac{\hat{e}_i x_{ji}}{\pi_i} - \frac{\hat{e}_t x_{jt}}{\pi_t}\right)\left(\frac{\hat{e}_i x_{ki}}{\pi_i} - \frac{\hat{e}_t x_{kt}}{\pi_t}\right)$$

where $\widehat{e}_i = y_i - \mathbf{x}_i \widehat{\mathbf{B}}_\pi$.

A $100 \times (1-\alpha)\%$ confidence interval for B_j can be worked out using the formula

$$\widehat{B}_{wj} \pm t_{\alpha/2, n-p-1}\sqrt{\widehat{V}\left(\widehat{B}_{wj}\right)} \text{ for } j = 0, 1, .., p \tag{20.2.20}$$

where $t_{\alpha/2, n-p-1}$ is the upper $\alpha/2 \times 100$ point of t-distribution with $n-p-1$ df.

20.2.2 Logistic Regression

In linear regression, the dependent variable y is generally taken as a continuous variable while in logistic regression the dependent variable y is binary taking values 1 if it possesses some attribute and 0 otherwise. For example, the HIV status (y) of person may be denoted as 1 if the person is HIV positive and 0 otherwise. We generally establish relationship between HIV status with age, race, occupation, economic condition, gender, etc. through a logistic regression model. The logistic regression with a vector of independent variable $\mathbf{x}' = (x_0, x_1, \ldots, x_p)$, $x_0 = 1$ and a parameter $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ is defined as

$$Prob(y = 1) = p(\mathbf{x}) = \frac{e^{\beta_0 + x_1\beta_1 + \cdots + x_p\beta_p}}{1 + e^{\beta_0 + x_1\beta_1 + \cdots + x_p\beta_p}} = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}} \qquad (20.2.21)$$

where $\mathbf{x}' = (1, x_1, \ldots, x_p)$ and $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$.

The probability y will take value zero is $Prob(y = 0) = 1 - p(\mathbf{x}) = \frac{1}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}}$. The odd ratio $p(\mathbf{x})/\{1 - p(\mathbf{x})\}$ is called log it$\{p(\mathbf{x})\}$ and it is linear in $\mathbf{x}$, i.e., log it$\{p(\mathbf{x})\} = \mathbf{x}'\boldsymbol{\beta}$. If the entire population of N units is surveyed, we would get the likelihood function assuming y_i's are independent as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1 - y_i} \qquad (20.2.22)$$

where $p_i = Prob(y_i = 1) = p(\mathbf{x}_i)$; y_i and $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})'$ are the values of y and $\mathbf{x}$ associated with the ith unit.

The likelihood equation $\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$ yields

$$\sum_{i=1}^{N}\left(y_i - \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)x_{ij} = 0 \quad \text{for} \quad j = 0, 1, \ldots p \qquad (20.2.23)$$

The finite population parameter $\mathbf{B} = (B_0, B_1, \ldots, B_p)'$ is defined as the solution for $\mathbf{B}$ of the equation

$$\sum_{i=1}^{N}\left(y_i - \frac{e^{\mathbf{x}_i'\mathbf{B}}}{1 + e^{\mathbf{x}_i'\mathbf{B}}}\right)x_{ij} = 0 \quad \text{for} \quad j = 0, 1, \ldots p \qquad (20.2.24)$$

No explicit expression of $\mathbf{B}$ can be obtained from Eq. (20.2.24). A design-based estimate of $\mathbf{B}$ is obtained as a solution of $\widehat{\mathbf{B}}_w$ of the equation

$$\sum_{i \in s} b_{si}\left(y_i - \frac{e^{\mathbf{x}_i'\hat{\mathbf{B}}_w}}{1 + e^{\mathbf{x}_i'\hat{\mathbf{B}}_w}}\right)x_{ij} = 0 \quad \text{for} \quad j = 0, 1, \ldots p \qquad (20.2.25)$$

where b_{si}'s are appropriately chosen weights. The solution of $\widehat{\mathbf{B}}_w$ may be obtained by Newton–Raphson iterative procedure. The variance of $\widehat{\mathbf{B}}_w$ can be obtained following the method described in Section 20.2.1. However, $\widehat{\boldsymbol{\beta}}$, the estimator of the model parameter $\boldsymbol{\beta}$ of the model (20.2.21) is obtained from the equation

$$\sum_{i\in s}\left(y_i - \frac{e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}\right)x_{ij} = 0 \quad \text{for} \quad j = 0, 1, \ldots p \tag{20.2.26}$$

20.3 MODEL-BASED APPROACH

In the model-based approach, the finite population is a realization of a random vector $\mathbf{Y} = (Y_1,\ldots, Y_N)$ under a superpopulation model ξ (see Chapter 6). Under this approach, we assume that the study variable y_i is related to the vector of auxiliary variables $\mathbf{x}_i$ through the following superpopulation model:

$$y_i = \boldsymbol{\beta}\mathbf{x}_i + \in_i \text{ for } i = 1, \ldots, N \tag{20.3.1}$$

where $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)'$; $E_\xi(\in_i) = 0$, $V_\xi(\in_i) = \sigma_i^2$ and $C_\xi(\in_i, \in_j) = 0$ for $i \neq j$; and E_ξ, V_ξ and C_ξ denote expectation, variance, and covariance with respect to the model ξ.

In case the entire population is surveyed and $\boldsymbol{\Sigma} = \text{diag}\left(\sigma_1^2, \ldots, \sigma_N^2\right)$ is known, one would get the GLS (generalized least square) estimator

$$\mathbf{B}^* = \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{20.3.2}$$

which is the parameter of the present interest.

In practice, surveying the entire population is not possible; hence a suitable estimator for $\mathbf{B}^*$ based on a sample s is the GLS estimator of $\boldsymbol{\beta}$, which is given by

$$\widehat{\boldsymbol{\beta}}_{gls} = \left(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{y}_s \tag{20.3.3}$$

where $\boldsymbol{\Sigma}_s$ denotes an $n \times n$ submatrix of $\boldsymbol{\Sigma}$ associated with the selected sample s, $\mathbf{X}_s$, and y_s are as defined in Section 20.2.

Obviously, $\widehat{\boldsymbol{\beta}}_{gls}$ is useful if $\boldsymbol{\Sigma}_s$ is known. In case $\sigma_i^2 = \sigma^2$ for $i = 1,\ldots, N$

$$\text{i.e.,} \quad \boldsymbol{\Sigma} = \sigma^2\ \mathbf{I}_N \tag{20.3.4}$$

the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is the OLS estimator

$$\widehat{\boldsymbol{\beta}}_{ols} = \left(\mathbf{X}_s'\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathrm{y}_s \tag{20.3.5}$$

The estimator $\widehat{\mathbf{B}}_\pi = \widehat{\boldsymbol{\beta}}_{gls}$ if $\pi_i \propto \sigma_i^2$ and $\widehat{\boldsymbol{\beta}}_{gls} = \widehat{\boldsymbol{\beta}}_{ols}$ if $\sigma_i^2 = \sigma^2$ for $i = 1, \ldots, N$.

20.3.1 Performances of the Proposed Estimators

All the estimators $\widehat{\mathbf{B}}_w$, $\widehat{\mathbf{B}}_\pi$, $\widehat{\boldsymbol{\beta}}_{gls}$, and $\widehat{\boldsymbol{\beta}}_{ols}$ are model unbiased for $\boldsymbol{\beta}$ in the sense that

$$E_\xi\left(\widehat{\mathbf{B}}_w\right) = E_\xi\left(\widehat{\mathbf{B}}_\pi\right) = E_\xi\left(\widehat{\boldsymbol{\beta}}_{gls}\right) = E_\xi\left(\widehat{\boldsymbol{\beta}}_{ols}\right) = \boldsymbol{\beta} \tag{20.3.6}$$

The model variances of $\widehat{\boldsymbol{\beta}}_{gls}$ and $\widehat{\boldsymbol{\beta}}_{ols}$ under the models (20.3.1) and (20.3.4) are, respectively,

$$V_\xi\left(\widehat{\boldsymbol{\beta}}_{gls}\right) = \left(\mathbf{X}'_s \boldsymbol{\Sigma}_s^{-1} \mathbf{X}_s\right)^{-1} \text{ and } V_\xi\left(\widehat{\boldsymbol{\beta}}_{ols}\right) = \sigma^2 \left(\mathbf{X}'_s \mathbf{X}_s\right)^{-1} \tag{20.3.7}$$

The estimator $\widehat{\boldsymbol{\beta}}_{gls} = \left(\widehat{\beta}_{0,gls}, \widehat{\beta}_{1,gls}, \ldots, \widehat{\beta}_{p,gls}\right)'$ is the optimum in the class of linear model unbiased estimators for $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)'$ under the model (20.3.1) in the sense $V_\xi\left(\widehat{\beta}_{j,gls}\right) \le V_\xi\left(\widehat{\beta}_j^*\right)$ for every $j = 0, 1, \ldots, p$, where $\widehat{\beta}_j^*$ is any other linear model unbiased estimator of β_j. Similarly, $\widehat{\boldsymbol{\beta}}_{ols}$ is optimal under model (20.3.4). However, no definite conclusion can be reached if the estimators are restricted to the class of model/design unbiased estimators, where the estimators $\widehat{\beta}_j^*$ satisfy $E_p E_\xi\left(\widehat{\beta}_j^*\right) = \beta_j$ and E_p is the expectation with respect to the sampling design p.

20.3.2 Variance Estimation

In case $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Sigma}^*$, where $\boldsymbol{\Sigma}^*$ is known, unbiased estimators of $V_\xi\left(\widehat{\boldsymbol{\beta}}_{gls}\right)$ and $V_\xi\left(\widehat{\boldsymbol{\beta}}_{ols}\right)$ under the model (20.3.1) and (20.3.4) are given, respectively, by

$$\widehat{V}_\xi\left(\widehat{\boldsymbol{\beta}}_{gls}\right) = \widehat{\sigma}_1^2 \left(\mathbf{X}'_s \boldsymbol{\Sigma}_s^{*-1} \mathbf{X}_s\right)^{-1} \tag{20.3.8}$$

and

$$\widehat{V}_\xi\left(\widehat{\boldsymbol{\beta}}_{ols}\right) = \widehat{\sigma}_2^2 \left(\mathbf{X}'_s \mathbf{X}_s\right)^{-1} \tag{20.3.9}$$

where

$$\widehat{\sigma}_1^2 = \frac{1}{n-p-1}\left(\mathbf{y}_s - \mathbf{X}_s\widehat{\boldsymbol{\beta}}_{gls}\right)'\left(\mathbf{y}_s - \mathbf{X}_s\widehat{\boldsymbol{\beta}}_{gls}\right) \text{ and}$$
$$\widehat{\sigma}_2^2 = \frac{1}{n-p-1}\left(\mathbf{y}_s - \mathbf{X}_s\widehat{\boldsymbol{\beta}}_{ols}\right)'\left(\mathbf{y}_s - \mathbf{X}_s\widehat{\boldsymbol{\beta}}_{ols}\right) \tag{20.3.10}$$

The model variances of $\widehat{\mathbf{B}}_w$ and $\widehat{\mathbf{B}}_\pi$ under model (20.3.1) with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^*$ are given by

$$V_m\left(\widehat{\mathbf{B}}_w\right) = \sigma^2\left(\mathbf{X}_s'\mathbf{W}_s\mathbf{X}_s\right)^{-1}\left(\mathbf{X}_s'\mathbf{W}_s\boldsymbol{\Sigma}_s^*\mathbf{W}_s\mathbf{X}_s\right)\left(\mathbf{X}_s'\mathbf{W}_s\mathbf{X}_s\right)^{-1} \tag{20.3.11}$$

and

$$V_m\left(\widehat{\mathbf{B}}_\pi\right) = \sigma^2\left(\mathbf{X}_s'\mathbf{W}_\pi\mathbf{X}_s\right)^{-1}\left(\mathbf{X}_s'\mathbf{W}_\pi\boldsymbol{\Sigma}_s^*\mathbf{W}_\pi\mathbf{X}_s\right)\left(\mathbf{X}_s'\mathbf{W}_\pi\mathbf{X}_s\right)^{-1} \tag{20.3.12}$$

The values of $\widehat{\boldsymbol{\beta}}_{ols}$, $\widehat{\boldsymbol{\beta}}_{gls}$, and $\widehat{\boldsymbol{B}}_\pi$ and estimated variances of $\widehat{\boldsymbol{\beta}}_{ols}$ and $\widehat{\boldsymbol{\beta}}_{gls}$ may be available by using standard software packages (e.g., BMDP), but the variance formulas (20.3.11) and (20.3.12) cannot be obtained by standard statistical packages unless the design is self-weighting ($\mathbf{W}_s \propto \mathbf{I}_n$) and $\boldsymbol{\Sigma}_s^* = \mathbf{I}_n$, where $\mathbf{I}_n$ is an identity matrix of order n.

20.3.3 Multistage Sampling

Consider a finite population consisting of N first-stage units (fsu's) and the ith fsu consists of M_i second-stage units (ssu's), $i = 1,\ldots, N$. Let a sample s of n fsu's be selected from the population by some suitable sampling design. If the ith fsu is selected in s, a subsample s_i of m_i ssu's is selected from the M_i ssu's of the ith fsu by some sampling procedure. Let y_{ij} be the value of the study variable y for the jth ssu of the ith fsu, $j = 1,\ldots, M_i$; $i = 1,\ldots, N$. Let us suppose that the study variable y is related to the explanatory variables $x_1,\ldots, x_p$ through the following superpopulation model

$$y_{ij} = \beta_0 x_{ij0} + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \in_{ij} \tag{20.3.13}$$

where $x_{ij0} = 1$, β_0, $\beta_1,\ldots, \beta_p$ are unknown model parameters, $\in_{ij}{}'s$ are error components with

$$E_\xi(\in_{ij}) = 0,\ V_\xi(\in_{ij}) = \sigma^2,\ C_\xi(\in_{ij}, \in_{ik}) = \rho\sigma^2,\ C_\xi(\in_{ij}, \in_{i'k}) = 0 \text{ for } i \neq i';$$
$$j = 1, \ldots, M_i;\ i, i' = 1, \ldots, N$$

The model (20.3.13) indicates that any two units belonging to the same fsu are correlated with a common intracluster correlation ρ, while ssu's belonging to different fsu's are uncorrelated. This type of model (20.3.13) was considered by Fuller (1975), Campbell (1977) and Holt and Scott (1981). For simplicity, let us assume that the selected fsu's in the sample s consist of first n units (i.e., fsu's labeled 1,2,..., n) and the subsample s_1 comprises of first m_1 ssu's of the fsu labeled 1, s_2 comprises of the first m_2 ssu's from the fsu labeled 2, and so on. The model for the selected sample can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{20.3.14}$$

where $\mathbf{Y} = \begin{pmatrix} y_{11} \\ \cdot \\ y_{1m_1} \\ y_{21} \\ \cdot \\ y_{2m_2} \\ \cdot \\ y_{n1} \\ \cdot \\ y_{nm_n} \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_{111} & x_{112} & \cdot & x_{11p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1m_1 1} & x_{1m_1 2} & \cdot & x_{1m_1 p} \\ 1 & x_{211} & x_{212} & \cdot & x_{21p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{2m_2 1} & x_{2m_2 2} & \cdot & x_{2m_2 p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n11} & x_{n12} & \cdot & x_{n1p} \\ & & \cdot & & \\ 1 & x_{nm_n 1} & x_{nm_n 2} & \cdot & x_{nm_n p} \end{pmatrix}$,

$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \beta_p \end{pmatrix}$ and. $\boldsymbol{\varepsilon} = \begin{pmatrix} \in_{11} \\ \cdot \\ \in_{1m_1} \\ \in_{21} \\ \cdot \\ \in_{2m_2} \\ \cdot \\ \in_{n1} \\ \cdot \\ \in_{nm_n} \end{pmatrix}$

The error component $\boldsymbol{\varepsilon}$ has $E_\xi(\boldsymbol{\varepsilon})=0$ and $V_\xi(\boldsymbol{\varepsilon})=\sigma^2\mathbf{G}$, where $\mathbf{G}=Diag(\mathbf{G}_1,\ldots,\ \mathbf{G}_\mathbf{n})$ with $\mathbf{G}_i=[(1-\rho)\mathbf{I}_{mi}+\rho\mathbf{E}_{mi,mi}]$, $\mathbf{I}_{mi}=$ unit matrix of order m_i, and $\mathbf{E}_{mi,mi}$ is a $m_i\times m_i$ matrix of each element 1. In case ρ is known and $\mathbf{X}$ is of full rank, the BLUE of $\boldsymbol{\beta}$ is the GLS estimator

$$\widehat{\boldsymbol{\beta}}_{gls}=\left(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{G}^{-1}\mathbf{Y} \tag{20.3.15}$$

The OLS estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{ols}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \tag{20.3.16}$$

Both the estimators $\widehat{\boldsymbol{\beta}}_{gls}$ and $\widehat{\boldsymbol{\beta}}_{ols}$ are model unbiased for $\boldsymbol{\beta}$ in the sense $E_\xi\left(\widehat{\boldsymbol{\beta}}_{gls}\right)=E_\xi\left(\widehat{\boldsymbol{\beta}}_{ols}\right)=\boldsymbol{\beta}$. The estimator $\widehat{\boldsymbol{\beta}}_{gls}$ is the optimal in the sense $V_\xi\left(\widehat{\beta}_{j,gls}\right)\le V_\xi\left(\widehat{\beta}_{j,ols}\right)$ for $j=0,\ldots,p$, where $\widehat{\beta}'_{gls}=\left(\widehat{\beta}_{0,gls},\widehat{\beta}_{1,gls},\ldots,\widehat{\beta}_{p,gls}\right)$ and $\widehat{\boldsymbol{\beta}}'_{ols}=\left(\widehat{\beta}_{0,ols},\widehat{\beta}_{1,ols},\ldots,\widehat{\beta}_{p,ols}\right)$.

In application of the GLS procedure the correlation coefficient ρ must be known. If it is not known, one may use iterative procedure following Fuller and Battese (1973) or use a computer package such as SUPERCARP (Hidiroglou et al., 1980). Furthermore, to compute $\widehat{\boldsymbol{\beta}}_{gls}$, one needs to know which observation comes from which cluster, and this information is rarely available, especially from a secondary data. The main advantage of $\widehat{\boldsymbol{\beta}}_{ols}$ is that it is easy to compute and does not require such information. For small value of ρ, the loss of efficiency of using $\widehat{\boldsymbol{\beta}}_{ols}$ is not substantial (Scott and Holt, 1982). The variances of $\widehat{\boldsymbol{\beta}}_{gls}$ and $\widehat{\boldsymbol{\beta}}_{ols}$ under the model (20.3.13) are given, respectively, by

$$V_\xi\left(\widehat{\boldsymbol{\beta}}_{gls}\right)=\sigma^2\left(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\right)^{-1} \tag{20.3.17}$$

and

$$\begin{aligned}V_\xi\left(\widehat{\boldsymbol{\beta}}_{ols}\right)&=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{G}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\\&=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{D}\end{aligned} \tag{20.3.18}$$

where $\mathbf{D}=\left(\mathbf{X}'\mathbf{G}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

If we ignore the effect of clustering, i.e., treating the correlation $\rho=0$ and accept the $V_\xi\left(\widehat{\boldsymbol{\beta}}_{ols}\right)=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$, then we will be committing a serious mistake of evaluating $V_\xi\left(\widehat{\boldsymbol{\beta}}_{ols}\right)$ while ignoring the factor $\mathbf{D}$. Kish and Frankel (1974) termed the factor $\mathbf{D}$ as a design effect although $\mathbf{D}$ does not depend on a sampling design. Hence, Scott and Holt (1982) termed $\mathbf{D}$ as a model "misspecification" effect. The effect of ignoring the factor $\mathbf{D}$ has a serious effect in estimating confidence interval and hypothesis testing based on $\widehat{\boldsymbol{\beta}}_{ols}$. The usual estimator of error variance σ^2 based on OLS (ignoring $\mathbf{D}$) is

$$\widehat{\sigma}^2_{ols} = \mathbf{Y}'(\mathbf{1} - \mathbf{P_0})\mathbf{Y}/(n - p - 1) \tag{20.3.19}$$

where $\mathbf{P_0} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Scott and Holt (1982) showed that under the model (20.3.13)

$$E_\xi\left(\widehat{\sigma}^2_{ols}\right) = \sigma^2\frac{1 - \mathrm{trace}(\mathbf{D})}{n - p - 1} = \sigma^2\frac{1 - (p + 1)\overline{d}}{n - p - 1} \tag{20.3.20}$$

where $\overline{d}$ is the average diagonal element of $\mathbf{D}$.

It is seen in most practical situations $\overline{d}$ exceeds 1, hence $\widehat{\sigma}^2_{ols}$ underestimates σ^2, but the effect is negligible if the sample size is reasonably large.

20.3.4 Separate Regression for Each First-Stage Unit

Konijn (1962), Porter (1973), and Pfeffermann and Nathan (1981) considered a separate regression model for each of the first-stage units. In this model, the dependent variable y for the ith fsu is related to a single independent variable x through the model

$$y_{ij} = \beta^{(i)}x_{ij} + \in_{ij} \text{ for } i = 1, \ldots, N; j = 1, \ldots, M_i \tag{20.3.21}$$

Let $\mathbf{Y}_i = \begin{pmatrix} y_{i1} \\ \cdot \\ y_{im_i} \end{pmatrix}$ and $\mathbf{X}_i = \begin{pmatrix} x_{i1} \\ \cdot \\ x_{m_i 1} \end{pmatrix}$ denote, respectively, the vector of dependent and independent variables for the sampled m_i ssu's that belong to the ith selected fsu in s, $E_\xi(\in_{ij}) = 0$, $V_\xi(\in_{ij}) = \sigma_i^2$ and $C_\xi(\in_{ij}, \in_{i'j'}) = 0$ for $(i,j) \neq (i'j')$.

Konijn (1962) selected a sample s with inclusion probability $\pi_i = \sum_{s \supset i} p(s)$ for the ith fsu and proposed an unbiased estimator of the average regression coefficient $\overline{\beta} = \dfrac{\sum_{i=1}^{N} M_i\beta^{(i)}}{M_0}$ as

$$\widehat{\overline{\beta}} = \frac{\sum_{i \in s} M_i \widehat{\beta}^{(i)}_{ols}}{M_0} \tag{20.3.22}$$

where $M_0 = \sum_{i=1}^{N} M_i$ and $\widehat{\beta}_{ols}^{(i)} = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{Y}_i$, the OLS estimator of $\beta^{(i)}$. The estimator $\widehat{\overline{\beta}}$ is a model design unbiased estimator of $\overline{\beta}$ because it satisfies $E_pE_m\left(\widehat{\overline{\beta}}\right) = \overline{\beta}$. Porter (1973) considered estimating $\overline{\beta}^* = \sum_{i=1}^{N} \beta^{(i)} \Big/ N$. Pfeffermann and Nathan (1981) proposed an alternative approach where the regression parameters $\beta^{(i)}$'s have a joint distribution $m_\beta(\theta)$. Although this model involves unknown parameters θ, it generally improves inference by using relationship between the variables $\beta^{(i)\prime}s(i = 1,\ldots, N)$. The model is very popular in econometrics, and good details are given by Maddala (1977), Fay and Herriot (1979), Rubin (1976), etc. In the proposed Pfeffermann and Nathan's (1981) model

$$\beta^{(i)} = \beta + \nu_i \tag{20.3.23}$$

$$E_m(\nu_i) = 0,\ V_m(\nu_i) = \delta^2 \text{ and } C_m(\nu_i, \nu_j) = 0 \text{ for } i \neq j;\ i,j = 1, \ldots, N$$

Following Haitovsky (1973), Pfeffermann and Nathan (1981) derived the extended least square estimators of the individual coefficients $\beta^{(i)}$ assuming δ^2 and σ_i^2 are known, as

$$\widehat{\widehat{\beta}}^{(i)} = \begin{cases} \lambda_i \widehat{\beta}^{(i)} + (1-\lambda_i)\widehat{\beta} & \text{for } i \in s \\ \widehat{\beta} & \text{otherwise} \end{cases} \tag{20.3.24}$$

where $\widehat{\beta}^{(i)} = \sum_{j\in s_i} x_{ij}y_{ij} \Big/ \sum_{j\in s_i} x_{ij}^2$, $\widehat{\beta} = \sum_{i\in s} \lambda_i \widehat{\beta}^{(i)} \Big/ \sum_{i\in s} \lambda_i$, and

$$\lambda_i = \left(1 + \frac{\sigma_i^2}{\delta^2 \sum_{j\in s_i} x_{ij}^2}\right)^{-1}.$$

The estimator $\widehat{\widehat{\beta}}^{(i)}$ cannot be used in practice because the parameters δ^2 and σ_i^2 are generally unknown. Pfeffermann and Nathan (1981) proposed estimators for σ_i^2 $(i \in s)$ as

$$\widehat{\sigma}_i^2 = \sum_{j\in s_i} \left(y_{ij} - \widehat{\beta}^{(i)} x_{ij}\right)^2 \Big/ (m_i - 1) \tag{20.3.25}$$

Denoting $\widehat{\lambda}_i = \left(1 + \dfrac{\hat{\sigma}_i^2}{\delta^2 \sum_{j\in s_i} x_{ij}^2}\right)^{-1}$, Pfeffermann and Nathan (1981) proposed the estimator δ^2 as the largest solution of the equation

$$\frac{1}{(n-1)}\sum_{i\in s}\widehat{\lambda}_i\left(\widehat{\beta}^{(i)}-\widehat{\overline{\beta}}^{(\hat{\lambda})}\right)^2=\delta^2 \tag{20.3.26}$$

where $\widehat{\overline{\beta}}^{(\hat{\lambda})}=\sum_{i\in s}\widehat{\beta}^{(i)}\widehat{\lambda}_i\Big/\sum_{i\in s}\widehat{\lambda}_i$.

Pfeffermann and Nathan (1981) studied in detail the existence and uniqueness of a positive solution and proximity of that solution to the true variance. Pfeffermann and Smith (1985) extended the results for p independent variables and studied the properties of the regression estimators and their application in large-scale real-life survey data.

20.4 CONCLUDING REMARKS

In descriptive inference the population parameter is supposed to be a known function of finite population values such as the population mean, variance, and correlation coefficient. The descriptive inference may be model dependent or design based. Analytic inference is model based. Kish and Frankel (1974) considered regression analysis in a descriptive approach. Jönrup and Rennermalm (1976) and Shah et al. (1977) followed a similar approach. On the basis of empirical studies, Kish and Frankel (1974) showed that if clustering of the population is ignored and OLS is used for regression analysis, then the estimates of the variances of the regression coefficients underestimate the variance. The underestimation of variances has serious effects on interval estimation and testing of hypotheses of parameters. The standard software packages provide wrong estimates of standard errors. However, Konjin (1962), Fuller (1975), and Pfeffermann and Smith (1985) favor analytic approach where the finite population is regarded as sample from an infinite population.

20.5 EXERCISES

20.5.1 Let a sample of 10 households be selected from a locality of 50 households by PPSWR method using household size (z) as a measure of size variable. The following table gives the monthly household expenditure on food (y), monthly household income (x), and household size (z). Fit a finite population regression y on x with intercept assuming mean household size of the population is 3.5. Find 95% confidence intervals of the regression coefficients.

House holds	1	2	3	4	5	6	7	8	9	10
z	4	5	2	6	4	2	3	3	4	1
y(in \$)	2000	2500	1000	4500	3500	2000	3000	2500	2800	1500
x (in \$)	5000	4000	2000	5600	4500	2500	6000	4000	3200	5000

20.5.2 Consider the Exercise 20.5.1., and assume that y is related to x through the regression model $y_i = \beta_0 + \beta_1 x_i + \in_i$; $E_m(\in_i) = 0$, $V_m(\in_i) = \sigma^2 x_i$, and $C_m(\in_i, \in_j) = 0$ for $i \neq j$. Estimate regression coefficients by OLS and GLS methods.

20.5.3 Let a sample of size 10 be selected from a finite population of 50 using a complex survey design. The following table gives the values of the study variable y, auxiliary variables x_1 and x_2, and inclusion probabilities π of the selected units in the sample. Estimate the finite population regression of y on x_1 and x_2, and also estimate the multiple correlation of y on x_1 and x_2.

Units	1	2	3	4	5	6	7	8	9	10
y	20	25	32	60	40	20	35	40	40	20
x_1	15	10	25	40	30	10	25	15	25	10
x_2	25	10	20	60	20	25	30	20	25	15
π	0.20	0.10	0.25	0.15	0.15	0.20	0.10	0.25	0.15	0.10

20.5.4 Let a sample of n first-stage units be selected from a population of N first-stage units. From each of the selected first-stage units, a subsample of m second-stage units is selected by some suitable sampling procedure. Let y_{ij} and x_{ijk} be the value of the study variable y and auxiliary variable x_k for the jth second-stage units of the ith first-stage units, $k = 1,2,\ldots, p$. Assume y_{ij} is related to x_{ij} through the superpopulation model:

$$y_{ij} = \beta_0 x_{ij0} + \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp} + \in_{ij}$$

where $x_{ij0} = 1$, $\beta_0, \beta_1, \ldots, \beta_p$ are unknown model parameters, $\in'_{ij}s$ are error components with

$$E_\xi(\in_{ij}) = 0,\ V_\xi(\in_{ij}) = \sigma^2,\ C_\xi(\in_{ij}, \in_{ik}) = \rho\sigma^2,$$
$$C_\xi(\in_{ij}, \in_{i'k}) = 0 \text{ for } i \neq i'; j = 1, \ldots, M_i; i, i' = 1, \ldots, N$$

Let $\widehat{\boldsymbol{\beta}}_G$ and $\widehat{\boldsymbol{\beta}}_0$ be the GLS and OLS of $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_p\right)'$, respectively. Show that (i) $\widehat{\boldsymbol{\beta}}_0$ is unbiased estimator of $\boldsymbol{\beta}$,

(ii) $\dfrac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2} \le E = \dfrac{\mathrm{Var}\left(c'\hat{\beta}_G\right)}{\mathrm{Var}\left(c'\hat{\beta}_0\right)} \le 1$, where $\mathbf{c}$ is an arbitrary coefficient vector and $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ are eigenvalues of

$\mathbf{V} = \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \cdot & \cdot & \ldots & \cdot \\ \rho & \rho & \ldots & 1 \end{bmatrix}$, and (iii) the upper bound of the loss of efficiency of using $\widehat{\boldsymbol{\beta}}_0$ in place of $\widehat{\boldsymbol{\beta}}_G$ is

$1 - E = \left(1 + \dfrac{4(1-\rho)[1+(m-1)\rho]}{m^2\rho^2}\right)^{-1}$ (Scott and Holt, 1982).