CHAPTER 10

Regression Estimation

10.1 Introduction

Scholastic success of students can be expected to depend to some extent on their scores on aptitude and achievement tests. Agricultural yields increase with the acreage and soil fertility. Industrial production may depend on employee sizes. Energy production and consumption of countries usually increase with their Gross National Products (GNPs). Consumer purchases frequently decrease with an increase in prices. Employment rates and stock prices are very much influenced by interest and inflation rates and related economic variables. Household expenses are directly related to family size.

In all these illustrations, the latter type of variables (x) provide auxiliary, concomitant, or supplementary information on the major variable (y). As the ratio method, the regression procedure utilizes this additional information to estimate the population total and mean of y with increased precision. For further benefits, this approach can also be combined with stratification. P.S.R.S. Rao (1987) presents a summary of the ratio and regression methods of estimation.

10.2 The regression estimator

In the illustration of Section 9.1 on test scores, the math score (y) is positively correlated with the total score (x). If the supplementary variable (x) and the major variable (y) are positively or negatively correlated, $(y_i - \bar{Y})$ can be expected to be very close to a constant multiple of $(x_i - \bar{X})$ for each of the population units. Consequently, departure of the sample mean of n units from the population mean, $(\bar{y} - \bar{Y})$, can be expected to be a multiple of the corresponding difference for the supplementary variable, $(\bar{x} - \bar{X})$. As a result, the **regression estimator** for $\bar{Y}$

that can be considered is

$$\hat{\bar{Y}}_l = \bar{y} + \beta(\bar{X} - \bar{x}), \qquad (10.1)$$

where $\beta = S_{xy}/S_x^2 = \rho S_y/S_x$ is the population regression coefficient.

This estimator is unbiased for $\bar{Y}$ and its variance is given by

$$\begin{aligned} V(\hat{\bar{Y}}_l) &= V(\bar{y}) + \beta^2 V(\bar{x}) - 2\beta \text{ Cov}(\bar{y}, \bar{x}) \\ &= \frac{(1-f)}{n}(S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}). \end{aligned} \qquad (10.2)$$

Note that $(S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}) = (S_y^2 - \beta^2 S_x^2) = S_y^2(1-\rho^2) = S_e^2$, which is the error or residual mean square.

Clearly, $V(\hat{\bar{Y}}_l)$ becomes smaller than $V(\bar{y})$ as ρ becomes positively or negatively large. It is also smaller than $V(\hat{\bar{Y}}_R)$ in (9.4) unless β coincides with R, that is, the regression of y goes through the origin as described in Section 9.7.

Example 10.1. Heights and weights: For the regression estimator of the average weight with height as the supplementary variable, from Table 9.1, $S_e^2 = 187.12(1 - 0.62^2) = 115.19$. For a sample of five from the 15 units, $V(\hat{\bar{Y}}_l) = 10(115.19)/75 = 15.36$ and hence S.E.$(\hat{\bar{Y}}_l) = 3.92$, which is slightly smaller than the S.E. of 3.97 found in Example 9.2 for the ratio estimator.

10.3 Estimation from the sample

From a sample (x_i, y_i), i = 1, 2,..., n, the slope β can be estimated from $b = s_{xy}/s_x^2$. Now, the regression estimator for $\bar{Y}$ is given by

$$\hat{\bar{Y}}_l = \bar{y} + b(\bar{X} - \bar{x}) \qquad (10.3)$$

The sample regression coefficient is not unbiased for β and as a result the estimator in (10.3) is biased for $\bar{Y}$. The bias of b and hence that of (10.3) become negligible for large n. The variance of (10.3) now becomes approximately the same as (10.2), and it can be estimated from

$$v(\hat{\bar{Y}}_l) = \frac{(1-f)}{n} s_e^2, \qquad (10.4)$$

where s_e^2 is the residual mean square, which can be expressed as $(s_y^2 + b^2 s_x^2 - 2bs_{xy}) = (s_y^2 - b^2 s_x^2) = s_y^2 (1 - r^2)$. Note that $r = s_{xy}/s_x s_y$ is the sample correlation coefficient.

Example 10.2. Heights and weights: For the five sample units (4, 6, 9, 12, 15) considered in Examples 9.1 and 9.2, $\bar{x} = 66$, $\bar{y} = 158.6$, $s_y^2 = 314.8$, $s_x^2 = 24$, and $s_{xy} = 84.5$. Hence $b = 84.5/24 = 3.52$ and $\hat{\bar{Y}}_l = 158.6 + 3.52(67.73 - 66) = 164.69$. Since $r = 84.5/(4.9 \times 17.74) = 0.97$, $v(\hat{\bar{Y}}_l) = (2/15)(314.8)(1 - 0.97^2) = 2.48$, and hence S.E.$(\hat{\bar{Y}}_l) = 1.57$.

10.4 Classical linear regression

As in the case of the ratio method, the linear regression procedure provides a motivation for the estimator in (10.3). In this approach, the principal characteristic y and the supplementary variable x are known as the dependent and independent or fixed variables. The **linear regression model** is

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{10.5}$$

The mean of y_i at x_i is given by $E(y_i | x_i) = \alpha + \beta x_i$, where α is the intercept and β is the slope of this regression line.

When $V(\varepsilon_i | x_i) = \sigma^2$ and ε_i and ε_j are uncorrelated, the LS estimators of the slope and intercept are obtained by minimizing $\Sigma(y_i - \alpha - \beta x_i)^2$ with respect to α and β. From this procedure, the estimators of these coefficients are given by

$$b = s_{xy}/s_{x^2} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}. \tag{10.6}$$

These estimators are unbiased. The variance of b is given by $V(b) = \sigma^2/\Sigma(x_i - \bar{x})^2$. Noting that $\text{Cov}(b, \bar{y}) = 0$, $V(a) = [\sigma^2/n + \bar{x}^2\sigma^2/\Sigma(x_i - \bar{x})^2]$.

The residual SS, $\Sigma(y_i - a - bx_i)^2$ has $(n - 2)$ d.f. and it can be expressed as $\Sigma(y_i - \bar{y})^2 - b^2\Sigma(x_i - \bar{x})^2$ = total SS − regression SS. An unbiased estimator of σ^2 with $(n - 2)$ d.f. is given by the residual or error mean square, EMS $= \Sigma(y_i - a - bx_i)^2/(n - 2)$. Note that the residual mean square s_e^2 in (10.4) is not based on the model in (10.5) and it is obtained by dividing the residual sum of squares with $(n - 1)$.

The LS estimator for $E(y_i | x_i)$ and the predicted value of y_i at x_i are obtained from $(a + bx_i) = \bar{y} + b(x_i - \bar{x})$, which is the best linear unbiased estimator (**BLUE**). Predicting the N values y_i of a finite population, its mean is obtained from $\hat{\bar{Y}}_l = \bar{y} + b(\bar{X} - \bar{x})$, which is the same as (10.3).

For the regression of the systolic pressures (y_i) with weight as the supplementary variable (x_i), from Table 9.1, b = (0.92)(12.6)/13.68 = 0.8475 and a = 141.33 − 0.8475(155.53) = 9.52. The mean $E(y_i \mid x_i)$ is estimated from 9.52 + 0.8475x_i. The same equation provides the predicted values of y_i for specified x_i.

In general, if the regression of y_i on x_i is of the form in (10.5) and $V(\varepsilon_i \mid x_i) = \sigma^2 x_i^h = \sigma^2 / W_i$, the WLS estimators of α and β are obtained by fitting the regression of $W_i^{1/2} y_i$ on $\alpha W_i^{1/2} + \beta W_i^{1/2} x_i$, that is, by minimizing $\Sigma W_i(y_i - \alpha - \beta x_i)^2$. Let $\bar{x}_w = \Sigma W_i x_i / \Sigma W_i$ and $\bar{y}_w = \Sigma W_i y_i / \Sigma W_i$ denote the weighted means. The estimators of the slope and intercept are given by $b = \Sigma W_i(x_i - \bar{x}_w)(y_i - \bar{y}_w) / \Sigma W_i(x_i - \bar{x}_w)^2$ and $a = \bar{y}_w - b\bar{x}_w$. These estimators are unbiased and their variances are given by $V(b) = \sigma^2 / \Sigma W_i(x_i - \bar{x}_w)^2$ and $V(a) = \sigma^2[1/\Sigma W_i + \bar{x}_w^2 / \Sigma W_i(x_i - \bar{x}_w)^2]$. An unbiased estimator of σ^2 is given by $\hat{\sigma}^2 = \Sigma W_i((y_i - a - bx_i)^2 / (n-2))$, which has $(n - 2)$ d.f. Both b/S.E.(b) and a/S.E.(a) follow Student's t-distribution with $(n - 2)$ d.f., and they can be used for testing β = 0 and α = 0, respectively. The significance of the regression, that is, β = 0, can also be tested from F = regression MS/residual MS = $b^2 \Sigma W_i(x_i - \bar{x}_w)^2 / \hat{\sigma}^2$, which follows the F-distribution with 1 and $(n - 2)$ d.f.

The population mean $\bar{Y}$ can be estimated from $\hat{\bar{Y}}_w = \bar{y}_w + b(\bar{X} - \bar{x}_w)$. Its variance is given by $V(\hat{\bar{Y}}_w) = \sigma^2[1/\Sigma W_i + (\bar{X} - \bar{x}_w)^2 / \Sigma W_i(x_i - \bar{x}_w)^2]$.

A graph of y_i plotted against x_i and examination of the variance of y_i at different values of x_i will be helpful in deciding on the suitable weighted ratio or regression type of estimator.

For the model in (10.5), Royall (1970) and Cochran (1977, pp. 199–200) show that the model-based optimum linear estimator takes the same form as $\hat{\bar{Y}}_l$ in (10.3). Royall (1976, 1986) and Valliant (1987) consider the model-based approach for the variance estimation for two-stage cluster sampling. Isaki and Fuller (1982), Särndal et al. (1992), Casady and Valliant (1993), and Tam (1986, 1995), among others consider the model-based approach for the estimation with sample survey data. Shah et al. (1977) describe the inference on the regression models from the data collected from surveys.

To estimate the population total Y or the mean $\bar{Y}$, the LS and the related model-based approaches result in linear combinations of the sample observations y_i, in general of the form $\Sigma l_i y_i$. Starting more than 45 years ago, Godambe (1955, 1966) made substantial theoretical contributions for finding optimum estimators utilizing all the available information on the population units that can be obtained, for example, from the auxiliary variables, the method of sampling, and the order of appearance of the population units in the sample.

10.5 Difference between regression estimators

Chapters 3 and 9 examined the difference between the sample means and the ratio estimators for two population means. The corresponding differences of the regression estimators can also be examined.

Common supplementary variable

Let $d_{1i} = y_{1i} + b_1(\bar{X} - x_i)$ and $d_{2i} = y_{2i} + b_2(\bar{X} - x_i)$. The sample means obtained from these expressions, $\bar{d}_1$ and $\bar{d}_2$, are the same as the regression estimators $\hat{\bar{Y}}_{l1} = \bar{y}_1 + b_1(\bar{X} - \bar{x})$ and $\hat{\bar{Y}}_{l2} = \bar{y}_2 + b_2(\bar{X} - \bar{x})$. The difference $(\hat{\bar{Y}}_{l1} - \hat{\bar{Y}}_{l2})$ is approximately unbiased for $(\bar{Y}_1 - \bar{Y}_2)$ and its variance can be estimated from

$$\begin{aligned} v(\hat{\bar{Y}}_{l1} - \hat{\bar{Y}}_{l2}) &= \frac{1-f}{n}\Sigma[(d_{1i} - d_{2i}) - (\bar{d}_1 - \bar{d}_2)]^2(n-1) \\ &= \frac{(1-f)}{n}\Sigma[(y_{1i} - \bar{y}_1) - (y_{2i} - \bar{y}_2) \\ &\quad -(b_1 - b_2)(x_i - \bar{x})]^2/(n-1). \qquad (10.7) \end{aligned}$$

Example 10.3. Systolic and diastolic pressures: With the data in Table 9.1, from the sample units (4, 6, 9, 12, 15), for the regression of systolic pressures on weight, $b_1 = 0.7195$ and hence $\hat{\bar{Y}}_{l1} = 150.21$. Similarly, for the diastolic pressures, $b_2 = 0.5051$ and $\hat{\bar{Y}}_{l2} = 94.21$. Thus, $(\hat{\bar{Y}}_{l1} - \hat{\bar{Y}}_{l2}) = 56$. From (10.7), the sample variance and S.E. of this estimate are 0.4702 and 0.6857.

Different supplementary variables

For this case, let $d_{1i} = y_{1i} + b_1(\bar{X}_1 - x_{1i})$ and $d_{2i} = y_{2i} + b_2(\bar{X}_2 - x_{2i})$. The corresponding sample means, $\bar{d}_1$ and $\bar{d}_2$ are the same as the regression estimators $\hat{\bar{Y}}_{l1} = \bar{y}_1 + b_1(\bar{X}_1 - \bar{x}_1)$ and $\hat{\bar{Y}}_{l2} = \bar{y}_2 + b_2(\bar{X}_2 - \bar{x}_2)$. The difference between these two estimators, $(\hat{\bar{Y}}_{l1} - \hat{\bar{Y}}_{l2})$, is approximately unbiased for $(\hat{\bar{Y}}_1 - \bar{Y}_2)$, and its variance can be estimated from

$$v(\hat{\bar{Y}}_{l1} - \hat{\bar{Y}}_{l2}) = [(1-f)/n]\Sigma(e_{1i} - e_{2i})^2/(n-1), \qquad (10.8)$$

where $e_{1i} = (y_{1i} - \bar{y}_1) - b_1(x_{1i} - \bar{x}_1)$ and $e_{2i} = (y_{2i} - \bar{y}_2) - b_2(x_{2i} - \bar{x}_2)$.

Example 10.4. Systolic pressure reduction with exercise: For the systolic pressures, as seen in Example 10.3, $\hat{\bar{Y}}_l = 150.21$. For the observations after the fitness program, $s_{xy} = 210.15$ and $b = 210.15/279.8 = 0.7511$. If the average weight after the program is $\bar{x}_2 = 150$, $\hat{\bar{Y}}_{l2} = 144.2 + 0.7511(150 - 151.4) = 143.15$. Thus, the difference of the regression estimators before and after the program is $150.21 - 143.15 = 7.06$. From (10.8), the estimate of the variance is 0.0459, and hence it has an S.E. of 0.2143.

10.6 Regression estimation vs. stratification

With the observations in Table T8 in the Appendix, it was found in Section 9.8 that for estimating the average wheat production in 1997, $V(\hat{\bar{Y}}_{\text{st}}) = 0.84$ for samples of size four from each of the two strata. For the ratio estimation with the 1995 observations providing the supplementary information, $V(\hat{\bar{Y}}_R) = 0.213$ for a sample of eight countries.

The observed data and the regression line $\hat{y}_i = 0.6264 + 1.0157x_i$ are presented in Figure 10.1. For this data, $S_e^2 = 2.6$ and for a sample of eight countries $V(\hat{\bar{Y}}_l) = (12/160)2.6 = 0.195$ and S.E.$(\hat{\bar{Y}}_l) = 0.44$.

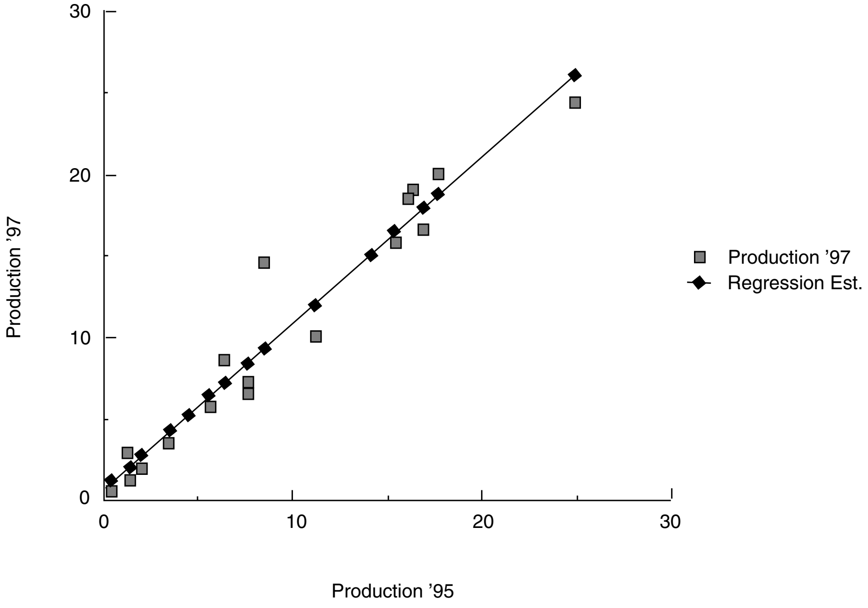

Figure 10.1. Regression of 1997 wheat production on 1995 production.

10.7 Stratification and regression estimator

Similar to the procedures in Section 9.9 for the ratio estimation, one can obtain the separate or combined regression estimators for the population total and mean.

Separate estimator

For the gth stratum, the regression estimate for $\bar{Y}_g$ is $\hat{\bar{Y}}_{gl} = \bar{y}_g + b_g(\bar{X}_g - \bar{x}_g)$, where b_g is the regression coefficient for the gth stratum. The separate estimator for the population mean is

$$\begin{aligned}\hat{\bar{Y}}_s &= (N_1\hat{\bar{Y}}_{1l} + N_2\hat{\bar{Y}}_{2l} + \cdots + N_G\hat{\bar{Y}}_{Gl})/N \\ &= W_1\hat{\bar{Y}}_{1l} + W_2\hat{\bar{Y}}_{2l} + \cdots + W_G\hat{\bar{Y}}_{Gl} = \Sigma W_g\hat{\bar{Y}}_{gl}. \qquad (10.9)\end{aligned}$$

Note that the numerator of the first expression is the regression estimator for the total.

For large n_g, the bias of this estimator will be negligible and its variance is approximately given by

$$V(\hat{\bar{Y}}_s) = \sum_{g=1}^{G} a_g S_{eg}^2, \qquad (10.10)$$

where $a_g = W_g^2(1 - f_g)/n_g$ and $S_{eg}^2 = S_{yg}^2(1 - \rho_g^2)$. For the estimate of this variance, replace S_{eg}^2 by $s_{eg}^2 = s_{yg}^2 - b_g^2 s_{xg}^2 = s_{yg}^2(1 - r_g^2)$.

For Neyman allocation of a sample of size n, the sizes n_g in the strata should be chosen proportional to $N_g S_{eg}$. If ρ_g are close to each other, n_g can be chosen proportional to $N_g S_{yg}$.

Example 10.5. Wheat production: For the first stratum considered in Section 9.7, the regression line is $\hat{y}_i = 0.4985 + 0.9591x_i$, with the S.E. of 0.46 and 0.099 for the intercept and slope. The corresponding t-statistics with $11 - 2 = 9$ d.f. are $t = 0.4985/0.46 = 1.1$ and $t = 0.9591/0.099 = 9.8$. From these figures, the intercept is not significantly different from zero at a 0.05 level of significance, but the slope is significantly different from zero at a level of significance of 0.005 or smaller.

For the second stratum, the regression line is $\hat{y}_i = 4.92 + 0.771x_i$, with the S.E. of 2.61 and 0.16 for the intercept and the slope. Judging from $t = 4.92/2.61 = 1.88$ and $t = 0.771/0.16 = 4.8$, each following the t-distribution with $9 - 2 = 7$ d.f., the intercept and slope are significantly different from zero at the levels of significance of 0.10 and 0.001, respectively.

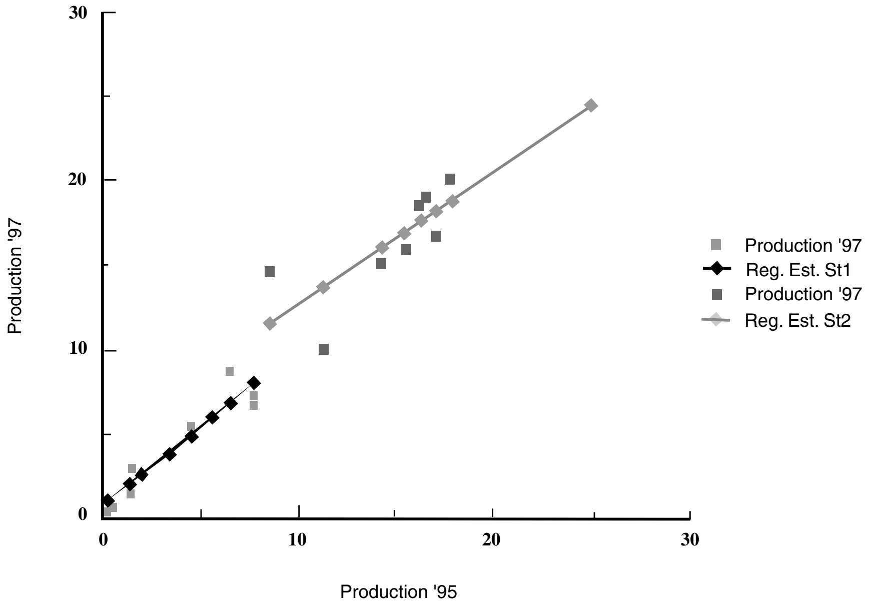

Figure 10.2. Regression of the wheat production for the two strata.

These regression lines are presented in Figure 10.2. From the data in Table T8 in the Appendix or the summary figures in Table 9.2, $s_{e1}^2 = 0.6386$ and $s_{e2}^2 = 3.5961$. Now, for samples of size four from the two strata, $v(\hat{\bar{Y}}_S) = (11/20)^2(7/44)(0.6386) + (9/20)^2\,(5/36)(3.5961) = 0.1319$, which is much smaller than the variance of 0.20 for the separate ratio estimator.

Combined regression estimator

If it is known that the slopes β_g in the strata do not differ much, one can consider the combined estimator:

$$\hat{\bar{Y}}_C = \Sigma W_g[\bar{Y}_g + \beta(\bar{X}_g - \bar{x}_g)] = \hat{\bar{Y}}_{\text{st}} + \beta(\bar{X} - \hat{\bar{X}}_{\text{st}}), \qquad (10.11)$$

where β is the common slope. The variance of this unbiased estimator is

$$V(\hat{\bar{Y}}_C) = \sum_{g=1}^{G} a_g(S_{yg}^2 + \beta^2 S_{xg}^2 - 2\beta S_{xyg}). \qquad (10.12)$$

This variance is minimized when $\beta = \Sigma t_g \beta_g / \Sigma t_g$, where $t_g = a_g S_{xg}^2$. With this optimum slope, the variance in (10.12) becomes

$$V\left(\hat{\bar{Y}}_C\right) = \sum a_g (S_{yg}^2 - \beta^2 S_{xg}^2). \tag{10.13}$$

From the samples in the strata, an estimator for β is $b = \Sigma t_g b_g / \Sigma t_g$. The estimator in (10.11) with β replaced by b is biased. For large samples, the bias will be small and the variance of $\hat{\bar{Y}}_C$ will be the same as (10.13).

With proportional allocation of the sample, b takes the form $\Sigma w_g s_{xyg} / \Sigma w_g s_{xg}^2$. The numerator and denominator of this expression are the weighted sum of cross-products and sum of squares, respectively. An estimate of the variance in (10.13) is

$$v(\hat{\bar{Y}}_C) = \sum a_g (s_{yg}^2 - b^2 s_{xg}^2). \tag{10.14}$$

For the two strata considered above, $a_1 = 0.048$, $a_2 = 0.028$, $t_1 = 0.137$, $t_2 = 0.128$. From these figures, $b = 0.8713$ and $v(\hat{\bar{Y}}_C) = 0.104$. This variance is smaller than the variance of 0.1319 for the separate regression estimator.

Preliminary analysis and tests of hypotheses

To examine whether the separate or combined estimator is appropriate, one can plot the observations of the entire population and of the strata, examine the variance of y_i at different values of x_i, and fit the suitable regressions as described at the end of Section 10.4. As a next step, one should test the hypotheses that the slopes and intercepts are significantly different from zero. As a final step, one should test the hypothesis that the regressions for the strata are not the same; that is, their slopes as well as intercepts are different.

The final step for the regressions of the wheat productions is next examined. The residual SS for the separate regressions is $\mathrm{SS}_{\mathrm{sep}} = 7.1 + 29.4 = 36.5$ with $9 + 7 = 16$ d.f. and the corresponding mean square is $\mathrm{MS}_{\mathrm{sep}} = 36.5/16 = 2.28$. For the regression with all the 20 observations, the residual SS is 49.49 with 18 d.f. The difference in the two SS is $\mathrm{SS}_{\mathrm{diff}} = 49.49 - 36.5 = 12.99$ with $18 - 16 = 2$ d.f. and the corresponding mean square $\mathrm{MS}_{\mathrm{diff}} = 12.99/2 = 6.45$. The ratio

$F = \mathrm{MS}_{\mathrm{diff}}/\mathrm{MS}_{\mathrm{sep}}$ follows the F-distribution with 2 and 16 d.f. For this illustration, $F = 6.45/2.28 = 2.83$. From the tables of this distribution, $F = 7.51$ for 2 and 16 d.f. when the significance level is 0.05. Thus, there is no significant difference between the regressions for the two strata, which suggests the combined regression estimator.

10.8 Multiple regression estimator

The regression method of estimation can include more than one supplementary variable. For example, one can estimate the average wheat production in 1997 utilizing the supplementary data from 1995 and 1990.

With two supplementary variables, x_1 and x_2, the population observations can be represented by (y_k, x_{1k}, x_{2k}), $k = 1, 2,..., N$. Denote the means of these three variables by $(\bar{Y}, \bar{X}_1, \bar{X}_2)$, variances by (S_{00}, S_{11}, S_{22}), and their covariances by (S_{01}, S_{02}, S_{12}). For a sample of size n, denote the means by $(\bar{y}, \bar{x}_1, \bar{x}_2)$, variances by (s_{00}, s_{11}, s_{22}), and their covariances by (s_{01}, s_{02}, s_{12}).

The multiple regression estimator for the population mean of $\bar{Y}$ is

$$\hat{\bar{Y}}_{Ml} = \bar{y} + \beta_1(\bar{X}_1 - \bar{x}_1) + \beta_2(\bar{X}_2 - \bar{x}_2), \tag{10.15}$$

where β_1 and β_2 are the population *regression coefficients* or *slopes*. This estimator is unbiased and its variance is minimized when $\beta_1 = (S_{22}S_{01} - S_{12}S_{02})/(S_{11}S_{22} - S_{12}^2)$ and $\beta_2 = (S_{22}S_{02} - S_{12}S_{01})/(S_{11}S_{22} - S_{12}^2)$. With these optimum values, the variance of (10.14) becomes

$$V(\hat{\bar{Y}}_{Ml}) = \frac{(1-f)}{n} S_e^2, \tag{10.16}$$

where $S_e^2 = \Sigma[(y_i - \bar{Y}) - \beta_1(x_{1i} - \bar{X}_1) - \beta_2(x_{2i} - \bar{X}_2)]^2/(N-1)$ is the residual mean square (MS).

From the sample, the estimators of β_1 and β_2 are given by $b_1 = (s_{22}s_{01} - s_{12}s_{02})/(s_{11}s_{22} - s_{12}^2)$ and $b_2 = (s_{22}s_{02} - s_{12}s_{01})/(s_{11}s_{22} - s_{12}^2)$. Estimating the slopes, (10.15) becomes

$$\hat{\bar{Y}}_{M1} = \bar{y} + b_1(\bar{X}_1 - \bar{x}_1) + b_2(\bar{X}_2 - \bar{x}_2). \tag{10.17}$$

This estimator, however, is not unbiased for the population mean. For large samples, its bias becomes negligible and its variance is approximately given by (10.16). An estimator of this variance is

$$v(\hat{\bar{Y}}_{Ml}) = \frac{(1-f)}{n} s_e^2, \tag{10.18}$$

where $s_e^2 = \Sigma[(y_i - \bar{y}) - b_1(x_{1i} - \bar{x}_1) - b_2(x_{2i} - \bar{x}_2)]^2/(n-1)$ is the sample residual MS.

The model for the classical multiple linear regression is

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \tag{10.19}$$

where β_1 and β_2 are the *slope* coefficients and α is the *intercept*. For given values x_{1i} and x_{2i} of the *independent variables*, the expectation of the residual ε_i is assumed to be zero; that is, $E(\varepsilon_i | x_{1i}, x_{2i}) = 0$, and its variance is assumed to be σ^2, the same at different values of the independent variables. For some applications, the last assumption for the variance may not be valid and it is suitably modified. Further, ε_i and ε_j are assumed to be uncorrelated. The expectation of the *dependent variable* is $E(y_i | x_{1i}, x_{2i}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}$, which is the equation to a plane.

From the least squares principle, the slopes are estimated from b_1 and b_2, and the intercept from $a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$. The estimator for the plane is given by $a + b_1 x_{1i} + b_2 x_{2i}$. The same expression is used for predicting y_i for specified (x_{1i}, x_{2i}). The residual mean square, $\Sigma[(y_i - \bar{y}) - b_1(x_{1i} - \bar{x}_1) - b_2(x_{2i} - \bar{x}_2)]^2/(n-3)$, is an unbiased estimator for σ^2.

The estimator in (10.15) is also obtained by considering the finite population as a sample from an infinite superpopulation, and predicting y_i with the above procedure. This approach provides another motivation for (10.15).

The following example examines whether it is advantageous to include both supplementary variables for estimation of the mean of the main characteristic.

Example 10.6. Wheat production: For the wheat production in 1997 (y) with the supplementary information from 1995 (x_1) and 1990 (x_2), from the data in Table T8 in the Appendix and Table 9.2, $a = 0.61$,

$b_1 = 0.869$, and $b_2 = 0.129$, with S.E. 0.60, 0.12, and 0.09, respectively. Each of the t-statistics $0.61/0.60 = 1.02$, $0.869/0.12 = 7.55$, and $0.129/0.09 = 1.43$ have $20 - 3 = 17$ d.f. From these results, only the slope coefficient for 1995 (x_1) is significantly different from zero.

From the above data, $S_e^2 = 2.3253$ and for a sample of eight countries, $V(\hat{\bar{Y}}_{Ml}) = 12(2.3253)/160 = 0.1744$ and S.E.$(\hat{\bar{Y}}_{Ml}) = 0.42$. This variance and S.E. are only a little different from the S.E. of 0.44 found in Section 10.6 for the regression estimator with the supplementary information from 1995. The above results from the classical tests of hypotheses also lead to the same conclusion.

As in Section 10.7, the multiple regression estimator can also be considered along with stratification.

10.9 Double sampling regression estimator

If $\bar{X}$ is not known, as described in Section 9.11 for the ratio estimator, the double-sampling regression estimator for $\bar{Y}$ is obtained from

$$\hat{\bar{Y}}_{ld} = \bar{y} + b(\bar{x}_1 - \bar{x}), \tag{10.20}$$

where $\bar{x}_1$ is the mean of the first sample of n_1 units and $(\bar{x}, \bar{y})$ as well as b are obtained from the second sample of n units selected from the first sample. Following Appendix A3, for large n, the bias of this estimator becomes negligible and its variance becomes

$$V(\hat{\bar{Y}}_{ld}) = (N - n_1)S_y^2/Nn_1 + (n_1 - n)S_y^2(1 - \rho^2)/n_1 n. \tag{10.21}$$

For the case of selecting the second sample independent of the first, P.S.R.S. Rao (1972) replaces $\bar{x}_1$ in (10.20) by $\bar{x}_a = \mathbf{a}\bar{x}_1 + (1 - \mathbf{a})\bar{x}$ or the mean of the distinct units $\bar{x}_v$, as described in Section 9.11 for the ratio estimator.

The regression estimator for the rotation surveys take the form of (10.20) or its modifications. Through the model in (10.5), Dorfman (1994) derives a model-based variance estimator for (10.20). P.S.R.S. Rao (1998b) presents a brief summary on the double sampling regression estimators.

10.10 Generalized regression and calibration estimators

When the n sample units are selected with probabilities ϕ_i, to estimate the population total Y, Cassel et al. (1976) suggest the **generalized regression estimator** (GREG)

$$\hat{Y}_{\text{greg}} = \Sigma(y_i/\phi_i) + b[X - \Sigma(x_i/\phi_i)], \tag{10.22}$$

where $b = \Sigma(x_i y_i/\phi_\mathrm{i})/\Sigma(x_i^2/\phi_i)$.

This estimator can be expressed as $\Sigma d_i y_i$. To relate ϕ_i with a supplementary characteristic, Deville and Sarndal (1992) consider the **calibration estimator** $\hat{Y}_c = \Sigma w_i y_i$. The weights w_i are found by minimizing $\Sigma[(w_i - d_i)^2/d_i]$ with the constraint $\Sigma w_i x_i = X$. Wright (1983) combines the features of both the GREG and calibration methods to estimate Y.

Exercises

10.1. *Project*. Consider the 20 samples of size three of Exercise 2.10 for regression estimation of the mean of the math scores with the verbal scores as the supplementary variable. Find the expectation and bias of (a) the slope b and (b) the regression estimator $\hat{\bar{Y}}_l$ for the mean of the math scores. (c) Compare the exact variance and MSE of $\hat{\bar{Y}}_l$ with the approximate variance in (10.2). (d) Find the expectation of the approximate variance in (10.4) and its bias for estimating the variance and MSE of $\hat{\bar{Y}}_l$.

10.2. From the five sample units (4, 6, 9, 12, 15) selected from Table 9.1, consider the regression estimation for the mean of the diastolic pressures. Find (a) the estimate, (b) its S.E. and (c) the 95% confidence limits for the mean.

10.3. With the sample units (4, 6, 9, 12, 15), as in Example 10.3, find (a) the difference of the regression estimates of the means of the systolic and diastolic pressures utilizing weight as the supplementary variable, (b) the S.E. for the difference of the estimates, and (c) the 95% confidence limits for the difference of the means.

10.4. For the sample units (4, 6, 9, 12, 15), consider the diastolic pressures (80, 92, 95, 80, 93) after the fitness program.

(a) Estimate the change in the diastolic pressures after the fitness program through the regression method and (b) find the 95% confidence limits for the change.

10.5. *Project*. From the data in Table T4 in the Appendix for the largest enrollment, for 1990 and 1995, for a sample of size four, (a) find the variance of the regression estimators for the average of the public enrollment with total enrollment as the supplementary variable, and (b) compare these variances with those found in Exercise 9.11 for ratio estimators.

10.6. Find the variance of the difference of the regression estimators in Exercise 10.5 and compare it with the variance of the difference of the sample means.

10.7. From the data in Table T5 in the Appendix for the smallest enrollment, for a sample of size four, (a) find the variance of the difference of the regression estimators for 1995 for public and private enrollment with total enrollment as the supplementary variable, and (b) compare it with the variance of the difference of the sample means.

10.8. Divide the data in Table 9.1 into two strata with the ranges 63 to 66 and 68 to 72 inches for the heights, and consider the corresponding heights, weights, and blood pressures. (a) Fit the regressions of weight on height in the two strata and perform the test of the hypothesis for their difference. (b) Irrespective of the inference from the test in (a), find the separate and combined regression estimators for the average weight, and compare their standard errors. (c) Examine whether or not the conclusion in (b) agrees with the inference in (a).

10.9. *Project*. As in Exercise 9.14, compute the separate and combined regression estimates in (10.9) and (10.11). (a) From the averages and variances of the estimates, find the biases and MSEs for the two procedures. (b) Compare the exact MSEs in (a) with the approximate expressions in (10.10) and (10.12).

10.10. From the data of the 25 countries in Table T8 in the Appendix for wheat production, fit the regressions of the production for 1997 (y) on 1995 (x_1), on 1990 (x_2), and on both x_1 and x_2. To estimate the total production in 1997 for the 25 countries with a sample of 10 countries, find the standard errors of the regression estimators with x_1, x_2 and (x_1,x_2) providing the supplementary information.

10.11. Based on the wheat production in 1997, the 25 countries in Table T8 in the Appendix can be divided into three strata consisting of the first 11, the next 9, and the remaining 5 countries. To estimate the total production in 1997 (y) with 1995 (x) as the supplementary variable, find the S.E. of the separate regression estimator for samples of 4, 4, and 2 countries from the three strata.