Statistics and Application 统计学与应用, 2022, 11(2), 248-261
Published Online April 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.112026

Hans汉斯

比率估计与回归估计抽样技术中五个非常实用的R函数

刘诗洋[1*]，张应应[1,2*#]

[1]重庆大学数学与统计学院统计与精算学系，重庆
[2]重庆大学分析数学与应用重庆市重点实验室，重庆

收稿日期：2022年3月12日；录用日期：2022年3月25日；发布日期：2022年4月11日

摘　要

比率估计与回归估计是利用辅助变量信息用以提高估计精度的非常重要的抽样技术。但在文献中，还没有方便的可以用于在仅给定基本的样本数据时得出总体均值与总体总值的比率估计量与回归估计量及其标准误差和置信区间的通用的R函数(程序)。本文自编了五个通用的R函数(程序)：Compute_R_ratio()、Compute_Y_bar_Y_MR()、Compute_Y_bar_Y_ratio()、Compute_Y_bar_Y_lr()及Compute_Y_bar_Y_Rs_Rc_lrs_lrc()，它们将会为需要使用比率估计及回归估计抽样技术以提高估计精度进行实际问题分析的使用者提供极大的方便。

关键词

比率估计，回归估计，分层估计，点估计和区间估计，R函数

Five Very Practical R Functions in Ratio Estimation and Regression Estimation Sampling Techniques

Shiyang Liu[1*], Yingying Zhang[1,2*#]

[1]Department of Statistics and Actuarial Science, College of Mathematics and Statistics, Chongqing University, Chongqing
[2]Chongqing Key Laboratory of Analytic Mathematics and Applications, Chongqing University, Chongqing

Received: Mar. 12th, 2022; accepted: Mar. 25th, 2022; published: Apr. 11th, 2022

[*]共一作者。
[#]通讯作者。

文章引用: 刘诗洋, 张应应. 比率估计与回归估计抽样技术中五个非常实用的 R 函数[J]. 统计学与应用, 2022, 11(2): 248-261. DOI: 10.12677/sa.2022.112026

Abstract

Ratio estimation and regression estimation are very important sampling techniques to improve estimation accuracy by using auxiliary variable information. However, in the literature, there is no general R function (program) which can be used to obtain the ratio estimator, regression estimator, standard error and confidence interval of the population mean and total value given only basic sample data. In this paper, we have written five general R functions (programs): Compute_R_ratio(), Compute_Y_bar_Y_MR(), Compute_Y_bar_Y_ratio(), Compute_Y_bar_Y_lr(), and Compute_Y_bar_Y_Rs_Rc_lrs_lrc(). The R functions (programs) will provide a great convenience for the users who need to use ratio estimation and regression estimation sampling techniques to improve the estimation accuracy and analyze practical problems.

Keywords

Ratio Estimation, Regression Estimation, Stratification Estimation, Point Estimation and Interval Estimation, R Function

1. 引言

比率估计[1] [2]与回归估计[3]-[8]是抽样技术[9]-[20]中极为重要的一部分内容。同时 R 软件[21] [22]作为统计学的常用编程工具，它具有完全免费、简洁高效、运行方便等优点。本文选取 R 软件对抽样调查中比率估计与回归估计的相关点估计和区间估计问题进行了程序实现。针对比率估计与回归估计以及相应的分层估计，本文自编了五个非常实用的 R 函数：Compute_R_ratio() (用于计算总体比率的点估计和区间估计)、Compute_Y_bar_Y_MR() (用于计算一元及二元辅助变量下总体均值与总体总值的比率点估计和区间估计)、Compute_Y_bar_Y_ratio() (用于计算总体均值及总体总值的比率点估计和区间估计)、Compute_Y_bar_Y_lr() (用于计算总体均值及总体总值的回归点估计和区间估计)及 Compute_Y_bar_Y_Rs_Rc_lrs_lrc() (用于计算总体均值及总体总值的分层比率和分层回归点估计和区间估计)。我们在对这五个 R 函数进行输入变量及输出变量的解释后给出了相应实际问题的 R 程序实现。我们相信，这五个 R 函数将会为需要使用比率估计及回归估计抽样技术以提高估计精度进行实际问题分析的使用者提供极大的方便。

2. 比率估计与回归估计中五个非常实用的 R 函数及应用举例

我们推荐比率估计与回归估计中五个非常实用的 R 函数。

R 函数 1：Compute_R_ratio()

对于比率估计，给定

$$\sum_{i=1}^{n} x_i,\ \sum_{i=1}^{n} y_i,\ \sum_{i=1}^{n} x_i^2,\ \sum_{i=1}^{n} y_i^2,\ \sum_{i=1}^{n} x_i y_i,\ n,\ \alpha$$

得到计算总体比率 R 的点估计和区间估计的 R 函数(程序) Compute_R_ratio()。由于正文版面的限制，

函数 Compute_R_ratio()的内容及输入输出的解释放在了补充材料中(下载链接：https://pan.baidu.com/s/1shGvwDATk7bQmnL97JpnDA?pwd=1234，提取码：1234)。

下面我们举一个例子来说明 Compute_R_ratio()的使用方法。

例 1 ([18]中例 5.2)在某地区抽取了由 33 个住户组成的简单随机样本，对每户调查两个指标：人口数(x_i)和每天用于食品支出的费用(y_i)，经计算得

$$\sum_{i=1}^{33} x_i = 123,\ \sum_{i=1}^{33} x_i^2 = 533,\ \sum_{i=1}^{33} y_i = 907.2,\ \sum_{i=1}^{33} y_i^2 = 28224,\ \sum_{i=1}^{33} x_i y_i = 3595.5$$

试对该地区平均每人每天用于食品的支出进行估计，并求其置信度为 95%的置信区间。

解：显然现在需要估计总体比率 R。可以计算：

$$\hat{R} = \frac{\sum_{i=1}^{33} y_i}{\sum_{i=1}^{33} x_i} = 7.37561$$

$$v_2\left(\hat{R}\right) = \frac{n}{(n-1)\left(\sum_{i=1}^{n} x_i\right)^2}\left(\sum_{i=1}^{n} y_i^2 + \hat{R}^2\sum_{i=1}^{n} x_i^2 - 2\hat{R}\sum_{i=1}^{n} y_i x_i\right) = 0.2849919$$

$$se\left(\hat{R}\right) = \sqrt{v_2\left(\hat{R}\right)} = 0.5338464$$

$$L = \hat{R} - t\cdot se\left(\hat{R}\right) = 6.32929,\ U = \hat{R} + t\cdot se\left(\hat{R}\right) = 8.421929$$

代入数据计算如下：

```
> rm(list = ls(all = TRUE))
> source("subfunctions.R")
> res = Compute_R_ratio(sum_x = 123, sum_y = 907.2, sum_x2 = 533, sum_y2 = 28224, sum_xy = 3595.5, n = 33, alpha = 0.05); res
      R_hat v_2_R_hat    se_R_hat           L           U
1 7.37561 0.2849919 0.5338464 6.32929 8.421929
```

因此，该地区人均每天食品支出 7.38 元，区间估计为[6.33, 8.42]元。

R 函数 2：Compute_Y_bar_Y_MR()

对于比率估计，给定：

$$\bar{X}_1,\ \bar{X}_2,\ \bar{y},\ \bar{x}_1,\ \bar{x}_2,\ s_y^2,\ s_{x_1}^2,\ s_{x_2}^2,\ s_{yx_1},\ s_{yx_2},\ n,\ N,\ \alpha$$

得到计算一元及二元辅助变量下总体均值 $\bar{Y}$ 与总体总值 Y 的比率估计量及其方差、标准误差及区间估计的 R 函数(程序) Compute_Y_bar_Y_MR()。由于正文版面的限制，函数 Compute_Y_bar_Y_MR()的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明 Compute_Y_bar_Y_MR()的使用方法。

例 2 ([18]中例 5.4)为精确估计某地区皮棉总产量，在该地区 301 个村庄中简单随机抽取了 18 个村庄，在调查皮棉产量 y_i 的同时记录了皮棉种植面积 x_{1i} 和良种比例 x_{2i}。该地区皮棉种植总面积为 $X_1 = 7450\ \text{hm}^2$，采用良种的平均比例为 $\bar{X}_2 = 40.10\%$。现以种植面积和良种比例为辅助变量对皮棉总产量分别做一元比率估计和二元比率估计，并比较二者估计精度的差异。对调查数据经过计算得到以下结果：

$$f=\frac{n}{N}=0.05980066,\ \frac{1-f}{n}=0.0522333,\ \bar{X}_2=40.10/100$$

$$\bar{y}=13.7967,\ \bar{x}_1=24.43899,\ \bar{x}_2=38.4444/100,\ s_y^2=35.4858$$

$$s_{x_1}^2=74.6789187,\ s_{x_2}^2=174.9671/10000,\ s_{yx_1}=42.26167,\ s_{yx_2}=46.5118/100$$

解：对于一元比率估计，只使用皮棉种植面积作为辅助变量时，可得：

$$\hat{Y}_{R_1}=\frac{\bar{y}}{\bar{x}_1}X_1=\hat{R}_1X_1=4205.796$$

$$v\left(\hat{Y}_{R_1}\right)=\frac{N^2(1-f)}{n}\left(s_y^2+\hat{R}_1^2s_{x_1}^2-2\hat{R}_1s_{yx_1}\right)=54751.7$$

$$se\left(\hat{Y}_{R_1}\right)=\sqrt{v\left(\hat{Y}_{R_1}\right)}=233.9908$$

$$L_{Y_{R_1}}=\hat{Y}_{R_1}-t\cdot se\left(\hat{Y}_{R_1}\right)=3747.183,\ U_{Y_{R_1}}=\hat{Y}_{R_1}+t\cdot se\left(\hat{Y}_{R_1}\right)=4664.41$$

对于一元比率估计，只使用良种比例作为辅助变量时，可得：

$$\hat{Y}_{R_2}=\frac{\bar{y}}{\bar{x}_2}X_2=\hat{R}_2X_2=4331.646$$

$$v\left(\hat{Y}_{R_2}\right)=\frac{N^2(1-f)}{n}\left(s_y^2+\hat{R}_2^2s_{x_2}^2-2\hat{R}_2s_{yx_2}\right)=116587.9$$

$$se\left(\hat{Y}_{R_2}\right)=\sqrt{v\left(\hat{Y}_{R_2}\right)}=341.4497$$

$$L_{Y_{R_2}}=\hat{Y}_{R_2}-t\cdot se\left(\hat{Y}_{R_2}\right)=3662.417,\ U_{Y_{R_2}}=\hat{Y}_{R_2}+t\cdot se\left(\hat{Y}_{R_2}\right)=5000.876$$

对于二元比率估计，即同时使用皮棉种植面积及良种比例做辅助变量，可得：

$$\hat{Y}_{MR}=W_1\hat{Y}_{R_1}+W_2\hat{Y}_{R_2}=-43250.28$$

$$v\left(\hat{Y}_{MR}\right)=\frac{N^2\left(v_{11}v_{22}-v_{12}^2\right)}{v_{11}+v_{22}-2v_{12}}=-11588534$$

$$se\left(\hat{Y}_{MR}\right)=\sqrt{v\left(\hat{Y}_{MR}\right)}=\text{NaN}$$

$$L_{Y_{MR}}=\hat{Y}_{MR}-t\cdot se\left(\hat{Y}_{MR}\right)=\text{NaN},\ U_{Y_{MR}}=\hat{Y}_{MR}+t\cdot se\left(\hat{Y}_{MR}\right)=\text{NaN}$$

代入本题数据进行计算得到：

> res_Y_bar_Y_MR = Compute_Y_bar_Y_MR(X1_bar = 7450 / 301, X2_bar = 40.10 / 100, y_bar = 13.7967, x1_bar = 24.43899, x2_bar = 38.4444 / 100, s2_y = 35.4858, s2_x1 = 74.6789187, s2_x2 = 174.9671 / 10000, s_y_x1 = 42.26167, s_y_x2 = 46.5118 / 100, n = 18, N = 301, alpha = 0.05); res_Y_bar_Y_MR

Warning message:

In sqrt(v_y_bar_MR) : 产生了 NaNs

$df_0_MR

	f	t	R1_hat	R2_hat	v_11	v_22
1	0.05980066	1.959964	0.5645364	35.88741	0.6043168	1.286828

```
          v_12        w_1         w_2
1 0.9451204 378.0842 -377.0842
$df_Y_bar_R1
   y_bar_R1 v_y_bar_R1 se_y_bar_R1 L_Y_bar_R1 U_Y_bar_R1
1 13.97275   0.6043168    0.7773781    12.44911    15.49638
$df_Y_R1
   Y_hat_R1 v_Y_hat_R1 se_Y_hat_R1    L_Y_R1   U_Y_R1
1 4205.796      54751.7      233.9908 3747.183 4664.41
$df_Y_bar_R2
   y_bar_R2 v_y_bar_R2 se_y_bar_R2 L_Y_bar_R2 U_Y_bar_R2
1 14.39085    1.286828     1.134384     12.1675     16.6142
$df_Y_R2
   Y_hat_R2 v_Y_hat_R2 se_Y_hat_R2    L_Y_R2    U_Y_R2
1 4331.646     116587.9      341.4497 3662.417 5000.876
$df_Y_bar_MR
    y_bar_MR v_y_bar_MR se_y_bar_MR L_Y_bar_MR U_Y_bar_MR
1 -143.6886   -127.9074          NaN         NaN         NaN
$df_Y_MR
    Y_hat_MR v_Y_hat_MR se_Y_hat_MR L_Y_MR U_Y_MR
1 -43250.28   -11588534          NaN     NaN     NaN
```

上述程序结果给出了使用一元及二元辅助变量下关于总体均值及总体总值的估计量及其方差、标准误差及置信区间的信息。对于本题所要研究的总体总值来看，只使用种植面积作为辅助变量时，我们得到该地区皮棉总产量 $\hat{Y}_{R_1}=4205.796$，抽样标准误差为 $se\left(\hat{Y}_{R_1}\right)=233.9908$，且皮棉总产量的区间估计为[3747.183, 4664.41]；只使用良种比例作为辅助变量时，我们得到该地区皮棉总产量为 $\hat{Y}_{R_2}=4331.646$，抽样标准误差为 $se\left(\hat{Y}_{R_2}\right)=341.4497$，且皮棉总产量的区间估计为[3662.417, 5000.876]。我们从分别使用种植面积及良种比例作为辅助变量做比率估计来看，使用种植面积做辅助变量情况下，总体总值的抽样标准误差较小，且具有更小的区间估计范围，估计效果优于使用良种比例作为辅助变量时的比率估计。二元比率估计情形下，我们可以看到程序结果计算得到的皮棉总产量及方差为负值，显然为不合理的结果。通过我们即将做的模拟实验得到二元比率估计下的标准误差均比一元情形下有更好的结果，这说明我们所编写的 R 函数 Compute_Y_bar_Y_MR()为正确函数，出现本题不合理的结果的可能原因是本题输入数据可能存在一定的偏差。

现考虑对二元辅助变量下的函数进行程序模拟检验其正确性。

```
> y_bar = 13.7967; x1_bar = 24.43899; x2_bar = 38.4444 / 100; s2_y = 35.4858; s2_x1 = 74.6789187; s2_x2 = 174.9671 / 10000
> ## set.seed(i), i = 1, 2, 3, 4, 5 结果类似。
> set.seed(5)
> y = sort(rnorm(n = 18, mean = y_bar, sd = sqrt(s2_y))); y
 [1]  0.786819  6.317740  7.360806  7.412230  8.787727  9.020503
 [7] 10.011796 10.205180 10.238508 10.984007 12.094348 12.858268
[13] 12.968760 14.214540 14.619410 21.109689 22.043323 23.991746
```

```
> y_bar = mean(y); y_bar
[1] 11.94586
> s2_y = var(y); s2_y
[1] 33.65602
> x1 = sort(rnorm(n = 18, mean = x1_bar, sd = sqrt(s2_x1))); x1
 [1] 17.06939 18.76069 21.90281 22.19772 26.52006 27.16903 30.54660
 [8] 31.51662 32.22094 32.57834 32.66220 34.02863 34.95683 36.69801
[15] 37.12467 37.22198 37.39094 43.58433
> x1_bar = mean(x1); x1_bar
[1] 30.7861
> s2_x1 = var(x1); s2_x1
[1] 51.64541
> x2 = sort(rnorm(n = 18, mean = x2_bar, sd = sqrt(s2_x2))); x2
 [1] 0.1198312 0.1513506 0.2509079 0.2671148 0.2783034 0.2886334
 [7] 0.3061590 0.3235651 0.3240513 0.3655805 0.3745791 0.3752891
[13] 0.4092755 0.4588122 0.5196322 0.5779954 0.5894784 0.6351937
> x2_bar = mean(x2); x2_bar
[1] 0.3675418
> s2_x2 = var(x2); s2_x2
[1] 0.02071014
> s_y_x1 = cov(y, x1); s_y_x1
[1] 37.4813
> s_y_x2 = cov(y, x2); s_y_x2
[1] 0.8125443
```

代入模拟数据进行计算得到：

```
> res_Y_bar_Y_MR = Compute_Y_bar_Y_MR(X1_bar = 7450 / 301, X2_bar = 40.10 / 100, y_bar,
x1_bar, x2_bar, s2_y, s2_x1, s2_x2, s_y_x1, s_y_x2, n = 18, N = 301, alpha = 0.05); res_Y_bar_Y_MR

$df_0_MR
          f         t      R1_hat      R2_hat         v_11         v_22
1 0.05980066 1.959964 0.3880276 32.50203 0.6447927 0.1418217
        v_12        w_1       w_2
1 0.300132 -0.8495305 1.849531
$df_Y_bar_R1
  y_bar_R1 v_y_bar_R1 se_y_bar_R1 L_Y_bar_R1 U_Y_bar_R1
1 9.604005  0.6447927   0.8029898   8.030174   11.17784
$df_Y_R1
  Y_hat_R1 v_Y_hat_R1 se_Y_hat_R1    L_Y_R1    U_Y_R1
1 2890.805   58418.86     241.6999 2417.082 3364.529
$df_Y_bar_R2
```

```
  y_bar_R2 v_y_bar_R2 se_y_bar_R2 L_Y_bar_R2 U_Y_bar_R2
1 13.03331  0.1418217   0.3765922   12.29521   13.77142
$df_Y_R2
  Y_hat_R2 v_Y_hat_R2 se_Y_hat_R2   L_Y_R2   U_Y_R2
1 3923.028   12849.19    113.3543 3700.857 4145.198
$df_Y_bar_MR
  y_bar_MR  v_y_bar_MR se_y_bar_MR L_Y_bar_MR U_Y_bar_MR
1 15.94662 0.007332276  0.08562871   15.77879   16.11445
$df_Y_MR
  Y_hat_MR v_Y_hat_MR se_Y_hat_MR   L_Y_MR   U_Y_MR
1 4799.932   664.3116    25.77424 4749.415 4850.448
```

由以上模拟实验结果得到二元比率估计下的标准误差均比一元情形下更好，这说明我们所编写的 R 函数 Compute_Y_bar_Y_MR()为正确函数。

R 函数 3：Compute_Y_bar_Y_ratio()

对于比率估计，给定

$$\boldsymbol{x}, \boldsymbol{y}, X, n, N, \alpha$$

得到计算总体均值 $\bar{Y}$ 及总体总值 Y 的比率估计量及其方差、标准误差及区间估计的 R 函数(程序)Compute_Y_bar_Y_ratio()，其中 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 为辅助变量的数据向量，$\boldsymbol{y}=(y_1,\cdots,y_n)$ 为调查变量的数据向量。由于正文版面的限制，函数 Compute_Y_bar_Y_ratio()的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明 Compute_Y_bar_Y_ratio()的使用方法。

例 3 ([18]中例 5.5)某地区有规模以下工业企业 127 个，共有固定资产价值 6794.5 万元，从中随机抽取 20 个企业调查工业产值 y_i 及固定资产价值 x_i，如表 1 所示。试通过比率估计得到该地区规模以下工业总产值 $\hat{Y}_R$ 及抽样标准误差 $se\left(\hat{Y}_R\right)$。

Table 1. Enterprise fixed assets value and industrial output value (Unit: 10000 yuan)
表 1. 企业固定资产价值及工业产值(单位：万元)

固定资产价值	工业产值	固定资产价值	工业产值
35	32.0	50	45.5
43	40.2	70	65.0
50	47.5	62	56.0
40	41.5	58	55.0
55	51.0	52	57.0
58	53.4	63	54.2
38	33.8	64	56.5
45	42.8	53	48.2
47	45.6	54	49.8
42	40.8	56	49.2

解：计算得到：

$$f=\frac{n}{N}=0.1574803,\ t=Z_{\alpha/2}=1.959964,\ \overline{X}=\frac{X}{N}=53.5$$

$$\overline{x}=51.75,\ \overline{y}=48.25,\ s_y^2=67.74684$$

$$s_x^2=88.51316,\ s_{yx}=73.41316,\ \hat{R}=\frac{\overline{y}}{\overline{x}}=0.9323671$$

$$\hat{Y}_R=N\overline{y}_R=N\frac{\overline{y}}{\overline{x}}\overline{X}=6334.969$$

$$v\left(\hat{Y}_R\right)=N^2v\left(\overline{y}_R\right)=\frac{N^2\left(1-f\right)}{n}\left(s_y^2+\hat{R}^2s_x^2-2\hat{R}s_{yx}\right)=5297.026$$

$$se\left(\hat{Y}_R\right)=\sqrt{v\left(\hat{Y}_R\right)}=72.78067$$

$$L_{Y_R}=\hat{Y}_R-t\cdot se\left(\hat{Y}_R\right)=6192.321,\ U_{Y_R}=\hat{Y}_R+t\cdot se\left(\hat{Y}_R\right)=6477.616$$

代入题中数据得到：

```
> n = 20; N = 127; alpha = 0.05; X = 6794.5
> x_vector = c(35, 43, 50, 40, 55, 58, 38, 45, 47, 42, 50, 70, 62, 58, 52, 63, 64, 53, 54, 56)
> y_vector = c(32.0, 40.2, 47.5, 41.5, 51.0, 53.4, 33.8, 42.8, 45.6, 40.8, 45.5, 65.0, 56.0, 55.0, 57.0, 54.2, 56.5, 48.2, 49.8, 49.2)
> res_Y_bar_Y_ratio = Compute_Y_bar_Y_ratio(x_vector, y_vector, X, n, N, alpha); res_Y_bar_Y_ratio
$df_0_ratio
                f        t X_bar x_bar y_bar      s2_y      s2_x      s_yx
1 0.1574803 1.959964   53.5 51.75 48.25 67.74684 88.51316 73.41316
      R_hat
1 0.9323671
$df_Y_bar_ratio
  y_bar_ratio v_y_bar_ratio se_y_bar_ratio L_Y_bar_ratio
1    49.88164     0.3284162     0.5730761      48.75843
  U_Y_bar_ratio
1      51.00485
$df_Y_ratio
  Y_hat_ratio v_Y_hat_ratio se_Y_hat_ratio L_Y_ratio U_Y_ratio
1    6334.969     5297.026      72.78067   6192.321   6477.616
```

从上述程序结果我们可以得到工业均产值和工业总产值的比率估计量及其方差、标准误差和置信区间的信息。对于本题感兴趣的工业总产值的比率估计而言，工业总产值的比率估计量为 $\hat{Y}_R=6334.969$ 万元，抽样标准误差为 $se\left(\hat{Y}_R\right)=72.78067$ 万元，工业总产值的比率估计的置信度为 95%的区间估计为 [6192.321, 6477.616] 万元。

R 函数 4：Compute_Y_bar_Y_lr()

对于回归估计，给定

$$\boldsymbol{x},\ \boldsymbol{y},\ X,\ n,\ N,\ \alpha$$

得到计算总体均值 $\bar{Y}$ 及总体总值 Y 的回归估计量及其方差、标准误差和区间估计的 R 函数(程序)Compute_Y_bar_Y_lr()，其中 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 为辅助变量的数据向量，$\boldsymbol{y}=(y_1,\cdots,y_n)$ 为调查变量的数据向量。由于正文版面的限制，函数 Compute_Y_bar_Y_lr()的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明 Compute_Y_bar_Y_lr()的使用方法。

例 4 (已知信息同本文例 3)试通过回归估计得到该地区规模以下工业总产值 $\hat{Y}_{lr}$ 及抽样标准误差 $se(\hat{Y}_{lr})$。

解：计算得到：

$$f=\frac{n}{N}=0.1574803,\ t=Z_{\alpha/2}=1.959964,\ \bar{X}=\frac{X}{N}=53.5$$

$$\bar{x}=51.75,\ \bar{y}=48.25,\ s_y^2=67.74684$$

$$s_x^2=88.51316,\ s_{yx}=73.41316,\ b=\frac{s_{yx}}{s_x^2}=0.8294039$$

$$\bar{y}_{lr}=\bar{y}+b(\bar{X}-\bar{x})=\bar{y}-b(\bar{x}-\bar{X})=49.70146$$

$$v(\bar{y}_{lr})=\frac{1-f}{n}\left(s_y^2-\frac{s_{yx}^2}{s_x^2}\right)=0.2888866,\ se(\bar{y}_{lr})=\sqrt{v(\bar{y}_{lr})}=0.5374818$$

$$L_{\bar{Y}_{lr}}=\bar{y}_{lr}-t\cdot se(\bar{y}_{lr})=48.64801,\ U_{\bar{Y}_{lr}}=\bar{y}_{lr}+t\cdot se(\bar{y}_{lr})=50.7549$$

$$\hat{Y}_{lr}=N\bar{y}_{lr}=6312.085$$

$$v(\hat{Y}_{lr})=N^2v(\bar{y}_{lr})=4659.453,\ se(\hat{Y}_{lr})=N\cdot se(\bar{y}_{lr})=68.26018$$

$$L_{Y_{lr}}=NL_{\bar{Y}_{lr}}=6178.298,\ U_{Y_{lr}}=NU_{\bar{Y}_{lr}}=6445.873$$

代入数据计算得到：

```
> n = 20; N = 127; alpha = 0.05; X = 6794.5
> x_vector = c(35, 43, 50, 40, 55, 58, 38, 45, 47, 42, 50, 70, 62, 58, 52, 63, 64, 53, 54, 56)
> y_vector = c(32.0, 40.2, 47.5, 41.5, 51.0, 53.4, 33.8, 42.8, 45.6, 40.8, 45.5, 65.0, 56.0, 55.0, 57.0, 54.2, 56.5, 48.2, 49.8, 49.2)
> res_Y_bar_Y_lr = Compute_Y_bar_Y_lr(x_vector, y_vector, X, n, N, alpha); res_Y_bar_Y_lr
$df_0_lr
          f        t X_bar x_bar y_bar     s2_y     s2_x     s_yx
1 0.1574803 1.959964  53.5 51.75 48.25 67.74684 88.51316 73.41316
          b
1 0.8294039
$df_Y_bar_lr
  y_bar_lr v_y_bar_lr se_y_bar_lr L_Y_bar_lr U_Y_bar_lr
1 49.70146  0.2888866   0.5374818   48.64801    50.7549
$df_Y_lr
  Y_hat_lr v_Y_hat_lr se_Y_hat_lr   L_Y_lr   U_Y_lr
1 6312.085   4659.453    68.26018 6178.298 6445.873
```

从上述程序结果我们可以得到工业均产值和工业总产值的回归估计量及其方差、标准误差和置信区

间的信息。对于本题感兴趣的工业总产值的回归估计而言，工业总产值的回归估计量为$\hat{Y}_{lr}=6312.085$万元，抽样标准误差为$se\left(\hat{Y}_{lr}\right)=68.26018$万元，工业总产值的回归估计的置信度为95%的区间估计为[6178.298, 6445.873]万元。

R函数5：Compute_Y_bar_Y_Rs_Rc_lrs_lrc()

对于分层比率估计和分层回归估计，给定

$$W_h,\ n_h,\ f_h,\ \bar{y}_h,\ \bar{x}_h,\ \bar{X}_h,\ s_{y_h}^2,\ s_{x_h}^2,\ s_{yx_h},\ N,\ \alpha$$

得到计算总体均值$\bar{Y}$及总体总值Y的分层比率估计量和分层回归估计量及其方差、标准误差和区间估计的R函数(程序)Compute_Y_bar_Y_Rs_Rc_lrs_lrc()。由于正文版面的限制，函数Compute_Y_bar_Y_Rs_Rc_lrs_lrc()的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明Compute_Y_bar_Y_Rs_Rc_lrs_lrc()的用法。

例5 ([18]中例5.6)某县有300个村，小麦播种面积为23434亩，全部村子按地势分为平原、丘陵和山区三种类型，各按10%的比例抽样来调查亩产量，得到表2中所示数据。其中$\bar{y}_h$为今年平均亩产，$\bar{x}_h$、$\bar{X}_h$是去年平均亩产，$\bar{y}_h$、$\bar{x}_h$是样本数据，$\bar{X}_h$是总体数据。现通过各种分层估计量来对今年全县的平均亩产进行估计。

Table 2. Wheat yield survey data
表2. 小麦产量调查数据

类型	N_h/个	W_h	n_h/个	$\bar{y}_h$/斤	$\bar{x}_h$/斤	$\bar{X}_h$/斤	$s_{y_h}^2$/斤^2	$s_{x_h}^2$/斤^2	s_{yx_h}/斤^2
平原	99	0.33	10	583	561	568	1809	1503	1643
丘陵	138	0.46	14	445	437	439	1990	1937	1948
山区	63	0.21	6	290	274	271	1989	1892	1936

解：从上表我们可以看到，小麦产量的相邻两年数据呈较高正相关性，因而在估计小麦的今年亩产时，辅助变量选择去年的小麦产量具有一定的合理性。此处抽样比为：$f_1=f_2=f_3=0.1$。

现对分别比率估计、联合比率估计、分别回归估计、联合回归估计的理论公式做出如下阐述。由于本例只涉及计算总体均值的分层估计量，因此结果部分仅显示总体均值的分层估计量及相应方差、标准误差和区间估计。

首先令

$$g_h=\frac{W_h^2\left(1-f_h\right)}{n_h},\ \left(g_1,g_2,g_3\right)=\left(0.00980100,0.01360286,0.00661500\right)$$

可以计算得到如下结果。

1) 分别比率估计：

$$\hat{R}_h=\frac{\bar{y}_h}{\bar{x}_h},\ \left(\hat{R}_1,\hat{R}_2,\hat{R}_3\right)=\left(1.039216,1.018307,1.058394\right)$$

$$\bar{y}_{Rs}=\sum_{h=1}^{L}W_h\bar{y}_{R_h}=\sum_{h=1}^{L}W_h\hat{R}_h\bar{X}_h=460.6606$$

$$v\left(\bar{y}_{Rs}\right)=\sum_{h=1}^{L}g_h\left(s_{y_h}^2+\hat{R}_h^2s_{x_h}^2-2\hat{R}_hs_{yx_h}\right)=0.663121$$

$$se\left(\bar{y}_{Rs}\right)=\sqrt{v\left(\bar{y}_{Rs}\right)}=0.8143224$$

$$L_{\bar{Y}_{Rs}}=\bar{y}_{Rs}-t\cdot se\left(\bar{y}_{Rs}\right)=459.0646,\ U_{\bar{Y}_{Rs}}=\bar{y}_{Rs}+t\cdot se\left(\bar{y}_{Rs}\right)=462.2567$$

2) 联合比率估计：

$$\bar{y}_{st}=\sum_{h=1}^{L}W_h\bar{y}_h=457.99,\ \bar{x}_{st}=\sum_{h=1}^{L}W_h\bar{x}_h=443.69$$

$$\hat{R}_c=\frac{\bar{y}_{st}}{\bar{x}_{st}}=1.03223,\ \bar{X}=\sum_{h=1}^{L}W_h\bar{X}_h=446.29$$

$$\bar{y}_{Rc}=\hat{R}_c\bar{X}=\frac{\bar{y}_{st}}{\bar{x}_{st}}\bar{X}=460.6738$$

$$v\left(\bar{y}_{Rc}\right)=\sum_{h=1}^{L}g_h\left(s_{y_h}^2+\hat{R}_c^2s_{x_h}^2-2\hat{R}_cs_{yx_h}\right)=0.6748678$$

$$se\left(\bar{y}_{Rc}\right)=\sqrt{v\left(\bar{y}_{Rc}\right)}=0.8215034$$

$$L_{\bar{Y}_{Rc}}=\bar{y}_{Rc}-t\cdot se\left(\bar{y}_{Rc}\right)=459.0637,\ U_{\bar{Y}_{Rc}}=\bar{y}_{Rc}+t\cdot se\left(\bar{y}_{Rc}\right)=462.2839$$

3) 分别回归估计：

$$b_h=\frac{s_{yx_h}}{s_{x_h}^2},\ \left(b_1,b_2,b_3\right)=\left(1.093147,1.005679,1.023256\right)$$

$$\bar{y}_{lrs}=\sum_{h=1}^{L}W_h\bar{y}_{lr_h}=\sum_{h=1}^{L}W_h\left[\bar{y}_h+b_h\left(\bar{X}_h-\bar{x}_h\right)\right]=460.7957$$

$$v\left(\bar{y}_{lrs}\right)=\sum_{h=1}^{L}g_h\left(s_{y_h}^2-\frac{s_{yx_h}^2}{s_{x_h}^2}\right)=0.6006202$$

$$se\left(\bar{y}_{lrs}\right)=\sqrt{v\left(\bar{y}_{lrs}\right)}=0.7749969$$

$$L_{\bar{Y}_{lrs}}=\bar{y}_{lrs}-t\cdot se\left(\bar{y}_{lrs}\right)=459.2768,\ U_{\bar{Y}_{lrs}}=\bar{y}_{lrs}+t\cdot se\left(\bar{y}_{lrs}\right)=462.3147$$

4) 联合回归估计：

$$b_c=\frac{\sum_{h=1}^{L}g_hs_{yx_h}}{\sum_{h=1}^{L}g_hs_{x_h}^2}=1.033825$$

$$\bar{y}_{lrc}=\bar{y}_{st}+b_c\left(\bar{X}-\bar{x}_{st}\right)=460.6779$$

$$v\left(\bar{y}_{lrc}\right)=\sum_{h=1}^{L}g_h\left(s_{y_h}^2+b_c^2s_{x_h}^2-2b_cs_{yx_h}\right)=0.6747315$$

$$se\left(\bar{y}_{lrc}\right)=\sqrt{v\left(\bar{y}_{lrc}\right)}=0.8214204$$

$$L_{\bar{Y}_{lrc}}=\bar{y}_{lrc}-t\cdot se\left(\bar{y}_{lrc}\right)=459.068,\ U_{\bar{Y}_{lrc}}=\bar{y}_{lrc}+t\cdot se\left(\bar{y}_{lrc}\right)=462.2879$$

代入数据计算可得：

```
> N_h = c(99, 138, 63); N = sum(N_h);
> W_h = c(0.33, 0.46, 0.21)
```

```
> n_h = c(10, 14, 6)
> f_h = c(0.1, 0.1, 0.1)
> y_bar_h = c(583, 445, 290)
> x_bar_h = c(561, 437, 274) ## 437
> X_bar_h = c(568, 439, 271)
> s2_y_h = c(1809, 1990, 1989)
> s2_x_h = c(1503, 1937, 1892) ## 1937
> s_yx_h = c(1643, 1948, 1936)
> alpha = 0.05
> res_Y_bar_Y = Compute_Y_bar_Y_Rs_Rc_lrs_lrc(W_h, n_h, f_h, y_bar_h, x_bar_h, X_bar_h, s2_y_h, s2_x_h, s_yx_h, N, alpha); res_Y_bar_Y
$t
[1] 1.959964
$g_h
[1] 0.00980100 0.01360286 0.00661500

$Rs
$Rs$R_hat_h
[1] 1.039216 1.018307 1.058394
$Rs$df_Y_bar_Rs
  y_bar_Rs v_y_bar_Rs se_y_bar_Rs L_Y_bar_Rs U_Y_bar_Rs
1 460.6606  0.663121  0.8143224  459.0646  462.2567
$Rs$df_Y_Rs
  Y_hat_Rs v_Y_hat_Rs se_Y_hat_Rs  L_Y_Rs U_Y_Rs
1 138198.2  59680.89  244.2967 137719.4 138677

$Rc
$Rc$df_0_Rc
  y_bar_st x_bar_st  Rc_hat  X_bar
1  457.99  443.69 1.03223 446.29
$Rc$df_Y_bar_Rc
  y_bar_Rc v_y_bar_Rc se_y_bar_Rc L_Y_bar_Rc U_Y_bar_Rc
1 460.6738 0.6748678  0.8215034  459.0637  462.2839
$Rc$df_Y_Rc
  Y_hat_Rc v_Y_hat_Rc se_Y_hat_Rc  L_Y_Rc  U_Y_Rc
1 138202.1  60738.1  246.451 137719.1 138685.2

$lrs
$lrs$b_h
[1] 1.093147 1.005679 1.023256
```

```
$lrs$y_bar_lr_h
[1] 590.6520 447.0114 286.9302
$lrs$df_Y_bar_lrs
  y_bar_lrs v_y_bar_lrs se_y_bar_lrs L_Y_bar_lrs U_Y_bar_lrs
1  460.7957   0.6006202   0.7749969   459.2768   462.3147
$lrs$df_Y_lrs
  Y_hat_lrs v_Y_hat_lrs se_Y_hat_lrs L_Y_lrs  U_Y_lrs
1  138238.7   54055.82     232.4991  137783 138694.4

$lrc
$lrc$df_0_lrc
  y_bar_st x_bar_st       bc  X_bar
1   457.99   443.69 1.033825 446.29
$lrc$df_Y_bar_lrc
  y_bar_lrc v_y_bar_lrc se_y_bar_lrc L_Y_bar_lrc U_Y_bar_lrc
1  460.6779   0.6747315   0.8214204   459.068   462.2879
$lrc$df_Y_lrc
  Y_hat_lrc v_Y_hat_lrc se_Y_hat_lrc  L_Y_lrc  U_Y_lrc
1  138203.4   60725.83     246.4261 137720.4 138686.4
```

从上述程序实现结果我们可以得到分层抽样下总体均值及总体总值的各分层估计量及其方差、标准误差和置信区间的信息。对于本题感兴趣的总体均值信息而言，在进行分别比率估计的情况下，全县平均亩产为 $\bar{y}_{Rs}=460.6606$ 斤，估计标准误差为 $se(\bar{y}_{Rs})=0.8143224$ 斤；在进行联合比率估计的情况下，全县平均亩产为 $\bar{y}_{Rc}=460.6738$ 斤，估计标准误差为 $se(\bar{y}_{Rc})=0.8215034$ 斤；在进行分别回归估计的情况下，全县平均亩产为 $\bar{y}_{lrs}=460.7957$ 斤，估计标准误差为 $se(\bar{y}_{lrs})=0.7749969$ 斤；在进行联合回归估计的情况下，全县平均亩产为 $\bar{y}_{lrc}=460.6779$ 斤，估计标准误差为 $se(\bar{y}_{lrc})=0.8214204$ 斤。

四种分层估计方法的总体均值的估计值和标准误差的估计值见表 3。从表中我们容易得到以下结论：对于比率估计和回归估计而言，总有分别估计量的估计精度优于联合估计量；对于分别估计和联合估计而言，总有回归估计量的估计精度优于比率估计量。

Table 3. The estimated value of the overall mean and the estimated value of the standard error of the four stratification estimation methods
表 3. 四种分层估计方法的总体均值的估计值和标准误差的估计值

	总体均值的估计值	标准误差的估计值
分别比率估计	460.6606	0.8143224
联合比率估计	460.6738	0.8215034
分别回归估计	460.7957	0.7749969
联合回归估计	460.6779	0.8214204

3. 总结

本文对抽样技术中的比率估计及回归估计给出了自编的五个非常实用的 R 函数(程序)：Compute_R_

ratio() (用于计算总体比率 R 的点估计和区间估计)、Compute_Y_bar_Y_MR() (用于计算一元及二元辅助变量下总体均值 $\bar{Y}$ 与总体总值 Y 的比率点估计和区间估计)、Compute_Y_bar_Y_ratio() (用于计算总体均值 $\bar{Y}$ 及总体总值 Y 的比率点估计和区间估计)、Compute_Y_bar_Y_lr() (用于计算总体均值 $\bar{Y}$ 及总体总值 Y 的回归点估计和区间估计)及 Compute_Y_bar_Y_Rs_Rc_lrs_lrc() (用于计算总体均值 $\bar{Y}$ 及总体总值 Y 的分层比率和分层回归点估计和区间估计)。这五个 R 函数(程序)很好地解决了在仅给定基本的样本数据时如何得出总体均值 $\bar{Y}$ 与总体总值 Y 的各类估计量及其标准误差和置信区间的问题，为需要使用比率估计及回归估计抽样技术进行实际问题分析的使用者提供了方便。

基金项目

本研究受教育部人文社会科学研究西部和边疆地区项目(20XJC910001)，国家社科基金西部项目(21XTJ001)和国家自然科学基金面上项目(72071019)支持。

参考文献

[1] 李金昌. 成数抽样的比率估计[J]. 统计研究, 1996, 13(5): 63-68.
[2] 刘建平, 陈光慧. 通过对辅助变量的线性转化来改进比率估计[J]. 统计研究, 2006, 23(7): 69-71.
[3] 钟守洋. 抽样回归估计方法的应用[J]. 统计与决策, 1986(1): 18-20+14.
[4] 袁卫, 刘文卿, 黎樟林. 中国农产量抽样调查的回归估计[J]. 统计研究, 1994(4): 37-39.
[5] 黄虎元. 轮换样本抽样下的回归估计[J]. 统计与信息论坛, 1997(3): 6-10.
[6] 金莹, 牛美玲, 汤银才. 整群抽样总体均值的回归估计[J]. 统计与决策, 2005(21): 4-5.
[7] 陈光慧. 多阶连续抽样设计下广义加权回归估计方法研究[J]. 数理统计与管理, 2019, 38(6): 996-1004.
[8] 杨贵军, 沈文静. 广义回归估计量在中国农业抽样调查中的应用研究[J]. 统计与信息论坛, 2020, 35(6): 10-16.
[9] 彤季. 抽样调查(V)——第四讲比估计与回归估计[J]. 数理统计与管理, 1987(5): 41-49.
[10] 邹国华, 周永正. 分层回归估计与分层比估计的方差性质[J]. 统计研究, 1996(5): 59-62.
[11] 冯士雍, 倪加勋, 邹国华. 抽样调查理论与方法[M]. 北京: 中国统计出版社, 1998.
[12] 冯士雍. 中国抽样调查应用中的若干问题[J]. 中国统计, 2001(11): 5-7.
[13] 金勇进, 蒋妍, 李序颖. 抽样技术[M]. 北京: 中国人民大学出版社, 2002.
[14] 俞纯权. 二阶抽样下的比估计与回归估计[J]. 统计与决策, 2006(1): 24-27.
[15] 侯延军. 抽样调查方法的妙用[J]. 调研世界, 2014(7): 62-63.
[16] 宛婉, 周国祥. Hadoop 平台的海量数据并行随机抽样[J]. 计算机工程与应用, 2014, 50(20): 115-118.
[17] 郭俊. 统计的魅力——漫谈统计思维、统计指标和抽样调查[J]. 调研世界, 2015(7): 61.
[18] 李金昌. 应用抽样技术[M]. 第三版. 北京: 科学出版社, 2015.
[19] 黄向阳. 大数据热潮下关于抽样的冷思考[J]. 中国统计, 2019(4): 51-54.
[20] 杨贵军, 尹剑, 孟杰, 王维真. 应用抽样技术[M]. 第二版. 北京: 中国统计出版社, 2020.
[21] 薛毅, 陈丽萍. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007.
[22] R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.