VLISS – Virtual Laboratory in Survey Sampling

Practical Methods for Design and Analysis of Complex Surveys.

Risto Lehtonen and Erkki Pahkinen

INSTRUCTIONS FOR TRAINING KEY 101a PART A: REGRESSION ESTIMATION (*Single Auxiliary Variable*)

Regression estimation of the total in the *Province'91* population. The previously selected SRSWOR sample is used. There, the study variable UE91 is regressed with the auxiliary variable HOU85. We conduct regression estimation in two ways, resulting in equal estimates. First we use SAS/SURVEYREG procedure and then we calculate the sum of the fitted values over the population. Both methods provide the desired regression estimate (also variance estimates in two ways). In Table 3.16, the sample identifiers correspond to the SRSWOR case, and the sampling rate is, as previously, 0.25.

Using UE91 as the dependent variable and HOU85 as the predictor, the slope is estimated as $\hat{b}$ = 0.152, giving: $\hat{t}_{reg} = \hat{t} + \hat{b}(T_z - \hat{t}_z) = 26440 + 0.152(91753 - 164952) = 15312$

Table 3.16 A simple random sample drawn without replacement from the *Province'91* population prepared for regression estimation.

Sample design identifiers			Element	Study var.	Auxiliary information			
					Variable	Model	WGHT	
STR	CLU	WGHT	LABEL	UE91	HOU85	Group	g-weight	w*-weight
1	1	4	Jyväskylä	4123	26881	1	0.2844	1.1378
1	4	4	Keuruu	760	4896	1	1.0085	4.0341
1	5	4	Saarijärvi	721	3730	1	1.0469	4.1877
1	15	4	Konginkangas	142	556	1	1.1057	4.6058
1	18	4	Kuhmoinen	187	1463	1	1.1216	4.4863
1	26	4	Pihtipudas	331	1946	1	1.1391	4.4227
1	30	4	Toivakka	127	834	1	1.1423	4.5691
1	31	4	Uurainen	219	932	1	1.1515	4.5562

Sampling rate = 8/32 = 0.25

The corresponding design-based total estimate obtained under SRSWOR was $\hat{t}$ = 26 440, whose standard error was 13 282. Therefore, the deff estimate is $deff = 648^2 / 13282^2 = 0.002$, which is almost zero and is persuasive evidence of the superiority of regression estimation over pure design-based estimation for the present estimation problem. Improved efficiency is due to the strong linear relationship between UE91 and HOU85.

An alternative method is based on the use of the calibrated weights w*. This will be demonstrated in Training Key 104.

VLISS Training key 101a: Yksi apumuuttuja (HOU85)

```
**Reference example 3.13;

**Regression estimation of the total in the Province'91 population. The previously
selected
SRSWOR sample is used. There, the study variable UE91 is regressed with the auxiliary
variable HOU85.
Regression estimation is conducted with the SAS/SURVEYREG procedure;

data Province91;
input Stratum Cluster Id Municipality $ 10-22 POP91 LAB91 UE91 HOU85 URB85;
datalines;
1  1   1  Jyväskylä    67200 33786  4123 26881  1
1  2   2  Jämsä        12907  6016   666  4663  1
 (...)
2  6  30  Toivakka      2499  1084   127   834  0
2  7  31  Uurainen      3004  1330   219   932  0
2  8  32  Viitasaari    8641  4011   568  3119  0
;
run;

**CODE 1: Frame Population";
proc print data=Province91;
title1 "Regression Estimation of Totals";
title2 "TABLE 2.1. Frame Population dataset Province91";
sum UE91 HOU85;
run;

**CODE 2: Population Correlation;
proc corr data=Province91;
title2 "Population correlation of UE91 with HOU85";
var UE91 HOU85;
run;

**CODE 3: Sample Selection;
data Sample;
set Province91;
SamplingWeight=4;
if id=1 or id=4 or id=5 or id=15 or id=18 or id=26 or id=30 or id=31 then output
Sample;
run;

proc print data=Sample;
title2 "TABLE 3.13. SRSWOR sample from Province91 population";
run;

**CODE 4: Sample Correlation;
proc corr data=Sample;
title2 "Sample correlation of UE91 with HOU85";
var UE91 HOU85;
run;

**CODE 5: Regression Estimation;
proc surveyreg data=Sample total=32;
title2 "Regression estimation for the total of UE91, auxiliary variable HOU85";
model UE91=HOU85 / solution;
weight SamplingWeight;
estimate "UE91 Total" Intercept 32 HOU85 91753 / E;
run;
```

Regression Estimation of Totals 1

TABLE 3.13. SRSWOR sample from Province91 population

Obs	Stratum	Cluster	Id	Municipality	POP91	LAB91	UE91	HOU85	URB85	Sampling Weight
1	1	1	1	Jyväskylä	67200	33786	4123	26881	1	4
2	1	2	4	Keuruu	12707	5919	760	4896	1	4
3	1	3	5	Saarijärvi	10774	4930	721	3730	1	4
4	2	3	15	Konginkangas	1636	675	142	556	0	4
5	2	2	18	Kuhmoinen	3357	1448	187	1463	0	4
6	2	8	26	Pihtipudas	5654	2543	331	1946	0	4
7	2	6	30	Toivakka	2499	1084	127	834	0	4
8	2	7	31	Uurainen	3004	1330	219	932	0	4

Sample correlation of UE91 with HOU85

The CORR Procedure

2 Variables: UE91 HOU85

Pearson Correlation Coefficients, N = 8

Prob > |r| under H0: Rho=0

	UE91	HOU85
UE91	1.00000	0.99912
		<.0001
HOU85	0.99912	1.00000
	<.0001	

Regression estimation for the total of UE91, auxiliary variable HOU85

The SURVEYREG Procedure

Regression Analysis for Dependent Variable UE91

Data Summary

Number of Observations	8
Sum of Weights	32.00000
Weighted Mean of UE91	826.25000
Weighted Sum of UE91	26440.0

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	42.6546808	22.1860968	1.92	0.0960
HOU85	0.1520142	0.0007745	196.29	<.0001

NOTE: The denominator degrees of freedom for the t tests is 7.

Coefficients of Estimate "UE91 Total"

Effect	Row 1
Intercept	32
HOU85	91753

Regression estimation for the total of UE91, auxiliary variable HOU85

Analysis of Estimable Functions

Parameter	Estimate	Standard Error	t Value	Pr > |t|
UE91 Total	15312.7108	648.160289	23.62	<.0001

INSTRUCTIONS FOR TRAINING KEY 101a PART A: REGRESSION ESTIMATION
(Multiple Regression Model)

Multiple regression estimation of the total in the Province'91 population. Here, the study variable UE91 is regressed with two auxiliary variables, HOU85 and a variable named URB85 with a value 1 for urban municipalities and zero otherwise . We calculate the estimates of the regression coefficients with SAS/SURVEYREG procedure and the regression estimated total first by the SURVEYREG procedure and then, by summing up the fitted values over the population.

We thus use both the formula (3.31) and the GREG method with equation (3.32). First, the estimated regression coefficients $\hat{b}_1$ and $\hat{b}_2$ are calculated by fitting a two-predictor regression model for the sample data set of *n* = 8 municipalities, as given in Table 3.13.

The estimates are $\hat{b}_1 = 0.14956$ and $\hat{b}_2 = 68.107$.

The estimated totals of auxiliary variables are $\hat{t}_{z_1} = 164952$, as previously, and $\hat{t}_{z_2} = 12$.

In addition, we use the known population totals $T_{z_1} = 91753$ and $T_{z_2} = 7$.

Using (3.31), we obtain:

$$\hat{t}_{reg} = \hat{t} + \hat{b}_1(T_{z_1} - \hat{t}_{z_1}) + \hat{b}_2(T_{z_2} - \hat{t}_{z_2}) = 26440+0.14956(91753\text{-}164952)+68.107(7\text{-}12)=15152.$$

Using (3.32), we obtain $\hat{t}_{reg} = \sum_{k=1}^{32} \hat{y}_k + \sum_{k=1}^{8} w_k(y_k - \hat{y}_k) = 15152 + 0 = 15152$. Thus, we first calculate the fitted values for all population elements. The sum of the fitted values over the population provides the desired regression estimate. The GREG estimation procedure is summarized in **Table 3.17** . There also, the estimate 15152 can be obtained. Note that in the SRSWOR case in which the sampling weights are equal to 4 and because there is an intercept in the regression model, the sum of the residuals over the sample data set is equal to zero. This is not the case if the weights would vary or if the intercept would be fixed to zero.

Calculating the multiple correlation coefficient squared $\hat{R}^2 = 0.998$ for the sample data set, we obtain the variance estimate of $\hat{t}_{reg}$ by (3.33), $\hat{v}(\hat{t}_{reg}) = 5692$, which is smaller than in the previous case where HOU85 was used as the only auxiliary variable. There, an estimate $\hat{v}(\hat{t}_{reg}) = 6482$ was obtained. Hence, multiple regression estimation appeared to be slightly more effective in this case. The design effect estimate is now $deff = 569^2 / 13282^2 = 0.0018$.

VLISS Training key 101a: Kaksi apumuuttujaa

```
**Reference example 3.13 (Two auxiliary variables (HOU85 and URB85));

**Multiple regression estimation of the total in the Province'91 population. The
study variable UE91 is
regressed with two auxiliary variables, HOU85 and URB85 (value 1 for urban
municipalities and zero for rural
municipalities). Calculation of the estimates of the regression coefficients and the
regression estimated total
with SAS/SURVEYREG procedure;

data Province91;
input Stratum Cluster Id Municipality $ 10-22 POP91 LAB91 UE91 HOU85 URB85;
datalines;
1  1   1  Jyväskylä    67200 33786  4123 26881  1
1  2   2  Jämsä        12907  6016   666  4663  1
(...)
2  6  30  Toivakka      2499  1084   127   834  0
2  7  31  Uurainen      3004  1330   219   932  0
2  8  32  Viitasaari    8641  4011   568  3119  0
;
run;

**CODE 1: Frame Population";
proc print data=Province91;
title1 "Regression Estimation of Totals";
title2 "TABLE 2.1. Frame Population dataset Province91";
sum UE91 HOU85 URB85;
run;

**CODE 2: Population Correlation;
proc corr data=Province91;
title2 "Population correlation of UE91 with HOU85 and URB85";
var UE91 HOU85 URB85;
run;

**CODE 3: Sample Selection;
data Sample;
set Province91;
SamplingWeight=4;
if id=1 or id=4 or id=5 or id=15 or id=18 or id=26 or id=30 or id=31 then output
Sample;
run;

proc print data=Sample;
title2 "TABLE 3.13. SRSWOR Sample from Province91 population";
run;

**CODE 4: Sample Correlation;
proc corr data=Sample;
title2 "Sample correlation of UE91 with HOU85 and URB85";
var UE91 HOU85 URB85;
run;

**CODE 5: Regression Estimation;
proc surveyreg data=Sample total=32;
title2 "Regression estimation for the total of UE91, auxiliary variables HOU85 and
URB85";
model UE91=HOU85 URB85/ solution;
weight SamplingWeight;
estimate "UE91 Total" Intercept 32 HOU85 91753 URB85 7/ E;
run;
```

Regression Estimation of Totals
TABLE 3.13. SRSWOR Sample from Province91 population

Obs	Stratum	Cluster	Id	Municipality	POP91	LAB91	UE91	HOU85	URB85	Sampling Weight
1	1	1	1	Jyväskylä	67200	33786	4123	26881	1	4
2	1	2	4	Keuruu	12707	5919	760	4896	1	4
3	1	3	5	Saarijärvi	10774	4930	721	3730	1	4
4	2	3	15	Konginkangas	1636	675	142	556	0	4
5	2	2	18	Kuhmoinen	3357	1448	187	1463	0	4
6	2	8	26	Pihtipudas	5654	2543	331	1946	0	4
7	2	6	30	Toivakka	2499	1084	127	834	0	4
8	2	7	31	Uurainen	3004	1330	219	932	0	4

Sample correlation of UE91 with HOU85 and URB85

The CORR Procedure
3 Variables: UE91 HOU85 URB85

Pearson Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	UE91	HOU85	URB85
UE91	1.00000	0.99912	0.63635
		<.0001	0.0898
HOU85	0.99912	1.00000	0.62092
	<.0001		0.1004

Regression estimation for the total of UE91, auxiliary variables HOU85 and URB85

The SURVEYREG Procedure
Regression Analysis for Dependent Variable UE91

Data Summary

Number of Observations	8
Sum of Weights	32.00000
Weighted Mean of UE91	826.25000
Weighted Sum of UE91	26440.0

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	29.7768913	19.7517828	1.51	0.1754
HOU85	0.1495578	0.0023199	64.47	<.0001
URB85	68.1072704	62.7319985	1.09	0.3136

Coefficients of Estimate "UE91 Total"

Effect	Row 1
Intercept	32
HOU85	91753
URB85	7

Analysis of Estimable Functions

Parameter	Estimate	Standard Error	t Value	Pr > |t|
UE91 Total	15151.9849	568.987386	26.63	<.0001

VLISS – Virtual Laboratory in Survey Sampling

Practical Methods for Design and Analysis of Complex Surveys.
Risto Lehtonen and Erkki Pahkinen

TRAINING KEY 104: Calibration of Weights

INSTRUCTIONS FOR TRAINING KEY 104: Calibration of Weights

---

1) The calibration equation for an auxiliary variable z in poststratification, ratio estimation and regression estimation is $\hat{t}_z = \sum_{k=1}^{n} w_k^* z_k = T_z$. This means that the HT estimator of the total of the auxiliary variable z reproduces the known population total of z-variable.

2) To verify this, we need to calculate:

a) Sampling weights $w_k = 1/\pi_k$
b) Adjustment weights (g-weights) g_k, which depend on both the chosen model-assisted technique and the realized sample. Formulas for different g-weights are given in page 89 (poststratification), page 93 (ratio estimation) and page 98 (regression estimation).
c) Calibrated weights $w^*_k = g_k w_k$.

• For ratio estimation and regression estimation the calibration property is verified by showing that the total $T_z = 91753$ will be reproduced.
• For poststratification the total $T_z = 7$ is reproduced being the number of urban municipalities in the population.

3) For poststratification and regression estimation, you can also make additional checks:

a) $\sum_{k=1}^{n} w_k = N$
b) $\sum_{k=1}^{n} g_k = n$
c) $\sum_{k=1}^{n} w_k^* = \sum_{k=1}^{n} g_k w_k = N$

4) For ratio estimation, the check 3a) only is valid (because the regression model used in ratio estimation does not include an intercept term).

We use a SRSWOR sample (n=8) selected from the Province'91 population to verify the calibration property. The results are shown first for poststratification, then for ratio estimation and finally for regression estimation. In addition, model-assisted estimates of the total of UE91 are calculated by using the calibrated weights.

Training Key 104: Calibration of weights
Calibration check
In the case of Ratio estimation

Obs	Id	LABEL	UE91	HOU85	SW	g_rat	wstar_rat	t_y_rat	t_z_rat
1	1	Jyväskylä	4123	26881	4	.5562	2.225	9173.52	59809.21
2	4	Keuruu	760	4896	4	.5562	2.225	1690.97	10893.42
3	5	Saarijärvi	721	3730	4	.5562	2.225	1604.20	8299.11
4	15	Konginkangas	142	556	4	.5562	2.225	315.94	1237.08
5	18	Kuhmoinen	187	1463	4	.5562	2.225	416.07	3255.12
6	26	Pihtipudas	331	1946	4	.5562	2.225	736.46	4329.78
7	30	Toivakka	127	834	4	.5562	2.225	282.57	1855.62
8	31	Uurainen	219	932	4	.5562	2.225	487.27	2073.66
			6610	**41238**	**32**	**4.450**	**17.80**	**14707.00**	**91753.00**

Calibration check
In the case of Regression estimation

Obs	Id	LABEL	UE91	HOU85	SW	g_reg	wstar_reg	t_y_reg	t_z_reg
1	1	Jyväskylä	4123	26881	4	.2845	1.138	4692.54	30594.26
2	4	Keuruu	760	4896	4	1.009	4.034	3065.90	19750.87
3	5	Saarijärvi	721	3730	4	1.047	4.188	3019.31	15620.02
4	15	Konginkangas	142	556	4	1.151	4.606	654.02	2560.81
5	18	Kuhmoinen	187	1463	4	1.122	4.486	838.94	6563.44
6	26	Pihtipudas	331	1946	4	1.106	4.423	1463.90	8606.51
7	30	Toivakka	127	834	4	1.142	4.569	580.28	3810.67
8	31	Uurainen	219	932	4	1.139	4.556	997.82	4246.41
			6610	**41238**	**32**	**8.000**	**32.00**	**15312.71**	**91753.00**