Multilevel Regression and Poststratification in Stata

Maurizio Pisati and Valeria Glorioso

Department of Sociology and Social Research
University of Milano-Bicocca (Italy)
maurizio.pisati@unimib.it   v.glorioso@campus.unimib.it

7th Italian Stata Users Group meeting
Bologna, November 11-12, 2010

Outline

1. Introduction
   - The problem
   - The solution

Outline

1. Introduction
   - The problem
   - The solution
2. Program

Outline

Outline

Outline

Introduction

A common research objective

- Sometimes social scientists are interested in determining whether, and to what extent, the distribution of a given variable of interest – which we will call the *criterion variable* and denote by symbol Y — varies across the categories of a second variable — which we will call the *discriminant variable* and denote by symbol D

A common research objective

- Sometimes social scientists are interested in determining whether, and to what extent, the distribution of a given variable of interest – which we will call the *criterion variable* and denote by symbol Y — varies across the categories of a second variable — which we will call the *discriminant variable* and denote by symbol D
- Without loss of generality, D can be taken to represent either a single categorical variable or the combination of two or more categorical variables

A common research objective

- The (conditional) distribution of Y within each category d of D can be described as follows:

$$Y_d \sim f(\theta_d, \phi_d) \quad \text{for } d = 1, \ldots, J$$

where $f(\cdot)$ denotes a generic probability distribution; θ_d denotes the expected value of the distribution; and ϕ_d denotes one or more ancillary parameters of the distribution (e.g., its variance)

A common research objective

- For the sake of simplicity, let us focus on the expected value of Y, so that our goal is to determine whether, and to what extent, the expected value of Y varies across the J categories of D

A common research objective

- For the sake of simplicity, let us focus on the expected value of Y, so that our goal is to determine whether, and to what extent, the expected value of Y varies across the J categories of D
- In terms of regression analysis, this amounts to estimating the J possible values of the regression function $E(Y|D = d)$, i.e., $E(Y|D = 1) \equiv \theta_1$, $E(Y|D = 2) \equiv \theta_2$, ..., $E(Y|D = J) \equiv \theta_J$

A common research objective

- For the sake of simplicity, let us focus on the expected value of Y, so that our goal is to determine whether, and to what extent, the expected value of Y varies across the J categories of D
- In terms of regression analysis, this amounts to estimating the J possible values of the regression function $E(Y|D = d)$, i.e., $E(Y|D = 1) \equiv \theta_1$, $E(Y|D = 2) \equiv \theta_2$, $\ldots$, $E(Y|D = J) \equiv \theta_J$
- Let us denote our estimand – i.e., our quantity of interest – by $\boldsymbol{\theta} \equiv \{\theta_d; d = 1, \ldots, J\}$

Estimating θ

- How do we get accurate – i.e., precise and unbiased – estimates of θ?

Estimating θ

- How do we get accurate – i.e., precise and unbiased – estimates of θ?
- For the sake of simplicity, let us suppose that (a) observations are sampled from a given target population, and (b) the data of interest are collected without measurement error, so that the only source of random estimation error is the sampling variance, and the only (possible) source of systematic estimation error is the selection bias

Estimating θ

- How do we get accurate – i.e., precise and unbiased – estimates of θ?
- For the sake of simplicity, let us suppose that (a) observations are sampled from a given target population, and (b) the data of interest are collected without measurement error, so that the only source of random estimation error is the sampling variance, and the only (possible) source of systematic estimation error is the selection bias
- The expression "selection bias" is used here as a shorthand for the sum of coverage bias, nonresponse bias, and sampling bias (Groves 1989)

Estimating $\boldsymbol{\theta}$

- The standard ML estimator of each element θ_d of $\boldsymbol{\theta}$ is:

$$\hat{\theta}_d \equiv \widehat{E(Y|D=d)} = \frac{\sum_{i=1}^{n_d} Y_i}{n_d}$$

where n_d denotes the number of valid sample observations within category d of variable D

Estimating θ

- When n_d is small, $\hat{\theta}_d$ tends to be very unprecise, i.e., to generate highly variable estimates of θ_d

Estimating θ

- When n_d is small, $\hat{\theta}_d$ tends to be very unprecise, i.e., to generate highly variable estimates of θ_d
- The accuracy of $\hat{\theta}_d$ decreases further if the data object of analysis are affected by selection bias, i.e., if the valid observations are a nonrandom sample of the target population *and* the process of selection into the sample is associated with one or more variables that are also associated with variable Y

Here's Mr. P

- For all those cases where the number of valid observations within one or more categories of D is small and/or collected data are affected by selection bias, relatively accurate estimates of θ can be obtained by using a proper combination of multilevel regression modeling and poststratification (henceforth MrP)

Here's Mr. P

- For all those cases where the number of valid observations within one or more categories of D is small and/or collected data are affected by selection bias, relatively accurate estimates of θ can be obtained by using a proper combination of multilevel regression modeling and poststratification (henceforth MrP)
- This approach has been devised by Andrew Gelman and colleagues (Gelman and Little 1997; Park, Gelman and Bafumi 2004; Park, Gelman and Bafumi 2006; Gelman and Hill 2007) and recently elaborated on by Kastellec, Lax and Phillips (Lax and Phillips 2009a; Lax and Phillips 2009b; Kastellec, Lax and Phillips 2010)

The MrP estimator

- The MrP estimator of θ – which we will denote by $\tilde{\theta}$ – can be described as a four-step procedure as follows:

The MrP estimator

- The MrP estimator of θ – which we will denote by $\tilde{\theta}$ – can be described as a four-step procedure as follows:

- **First:** Identify one or more variables that might possibly be responsible for selection bias. Without loss of generality, we will treat the full cross-classification of these variables as a single categorical variable, which we will denote by G

The MRP estimator

- **Second:** Define the new estimand $\boldsymbol{\gamma} \equiv \{\gamma_{d,g}; d = 1, \ldots, J; g = 1, \ldots, K\}$, where $\gamma_{d,g} \equiv E(Y|D = d, G = g)$; d indexes the J categories of variable D as above; and g indexes the K categories of variable G

The MrP estimator

- **Second:** Define the new estimand $\boldsymbol{\gamma} \equiv \{\gamma_{d,g}; d = 1, \ldots, J; g = 1, \ldots, K\}$, where $\gamma_{d,g} \equiv E(Y|D = d, G = g)$; d indexes the J categories of variable D as above; and g indexes the K categories of variable G
- **Third:** Use a properly specified multilevel regression model to estimate $\boldsymbol{\gamma}$

The MrP estimator

- **Fourth:** Compute the estimate of each element θ_d of $\boldsymbol{\theta}$ as a weighted sum of the proper subset of $\hat{\boldsymbol{\gamma}}$:

$$\tilde{\theta}_d = \sum_{g=1}^{G} \hat{\gamma}_{d,g} w_{g|d}$$

where $w_{g|d} = N_{g,d}/N_d$; N_d denotes the number of members of the target population who belong in category d of variable D; and $N_{g,d}$ denotes the number of members of the target population who belong in category d of variable D *and* in category g of variable G

The MrP estimator: Advantages

- The use of multilevel regression modeling (step 3 above) helps to increase precision

The MrP estimator: Advantages

- The use of multilevel regression modeling (step 3 above) helps to increase precision
- If variable G is carefully defined, poststratification (step 4 above) helps to decrease bias

The MrP estimator: Advantages

- The use of multilevel regression modeling (step 3 above) helps to increase precision
- If variable G is carefully defined, poststratification (step 4 above) helps to decrease bias
- In sum, we expect MrP to be a relatively accurate estimator of θ

The MrP estimator: Disadvantages

- We need to have population data for the full $D \times G$ cross-classification; this might limit the definition of G

The MrP estimator: Disadvantages

- We need to have population data for the full $D \times G$ cross-classification; this might limit the definition of G
- To get good estimates of $\boldsymbol{\gamma}$, the multilevel regression model must be specified very carefully – but this caveat applies to any kind of regression model

Program

Using mrp: An example

Example dataset

```
. describe

Contains data from /Users/Tonzolo/Lavori/MRP/Simul/xsamp.dta
  obs:         1,000
 vars:            10                          16 Oct 2010 09:13
 size:        17,000 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
region          byte   %21.0g      reg        Region of residence
area            byte   %17.0g      area       Area of residence
relmar          float  %4.1f                  Religious marriages (%)
sex             byte   %9.0g       sex        Sex
age             byte   %9.0g       age        Age
edu             byte   %21.0g      educ       Level of education
sex_age         byte   %12.0g      sex_age    Interaction sex*age
sex_edu         byte   %28.0g      sex_edu    Interaction sex*edu
age_edu         byte   %27.0g      age_edu    Interaction age*edu
church          byte   %9.0g       church     Church attendance
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

Using mrp: An example

Cross-tabulation $D \times Y$ - Absolute frequencies

```
. tab region church
```

Region of residence	Church attendance Irregular	Regular	Total
Piemonte	63	34	97
Lombardia	58	37	95
Trentino-Alto Adige	40	19	59
Veneto	33	29	62
Friuli-Venezia Giulia	30	9	39
Liguria	37	14	51
Emilia-Romagna	32	19	51
Toscana	47	11	58
Umbria	24	9	33
Marche	26	9	35
Lazio	39	25	64
Abruzzo	21	13	34
Molise	16	11	27
Campania	36	27	63
Puglia	35	26	61
Basilicata	19	7	26
Calabria	16	17	33
Sicilia	36	25	61
Sardegna	38	13	51
Total	646	354	1,000

Using mrp: An example

$E(Y|D = d)$ – Standard ML estimator and MrP estimator

```
. tab region church, row nofre
```

Region of residence	Church attendance Irregular	Regular	Total
Piemonte	64.95	35.05	100.00
Lombardia	61.05	38.95	100.00
Trentino-Alto Adige	67.80	32.20	100.00
Veneto	53.23	46.77	100.00
Friuli-Venezia Giulia	76.92	23.08	100.00
Liguria	72.55	27.45	100.00
Emilia-Romagna	62.75	37.25	100.00
Toscana	81.03	18.97	100.00
Umbria	72.73	27.27	100.00
Marche	74.29	25.71	100.00
Lazio	60.94	39.06	100.00
Abruzzo	61.76	38.24	100.00
Molise	59.26	40.74	100.00
Campania	57.14	42.86	100.00
Puglia	57.38	42.62	100.00
Basilicata	73.08	26.92	100.00
Calabria	48.48	51.52	100.00
Sicilia	59.02	40.98	100.00
Sardegna	74.51	25.49	100.00
Total	64.60	35.40	100.00

Using mrp: An example

$E(Y|D = d)$ – Standard ML estimator and MrP estimator

```
. tab region church, row nofre
```

Region of residence	Church attendance Irregular	Regular	Total
Piemonte	64.95	35.05	100.00
Lombardia	61.05	38.95	100.00
Trentino-Alto Adige	67.80	32.20	100.00
Veneto	53.23	46.77	100.00
Friuli-Venezia Giulia	76.92	23.08	100.00
Liguria	72.55	27.45	100.00
Emilia-Romagna	62.75	37.25	100.00
Toscana	81.03	18.97	100.00
Umbria	72.73	27.27	100.00
Marche	74.29	25.71	100.00
Lazio	60.94	39.06	100.00
Abruzzo	61.76	38.24	100.00
Molise	59.26	40.74	100.00
Campania	57.14	42.86	100.00
Puglia	57.38	42.62	100.00
Basilicata	73.08	26.92	100.00
Calabria	48.48	51.52	100.00
Sicilia	59.02	40.98	100.00
Sardegna	74.51	25.49	100.00
Total	64.60	35.40	100.00

```
. mrp church region using PostStrat.dta,                      ///
>     yvartype(categorical)                                   ///
>     group(sex age edu)                                      ///
>     psw(N)                                                  ///
>     model(linear)                                           ///
>     linpred(c.relmar R.area R.region R.sex_age R.sex_edu    ///
>             R.age_edu)                                      ///
>     percent tableopt(format(%5.1f) row)
```

Region of residence	Church attendance Irregular	Regular
Piemonte	66.7	33.3
Lombardia	63.1	36.9
Trentino-Alto Adige	66.8	33.2
Veneto	60.4	39.6
Friuli-Venezia Giulia	72.8	27.2
Liguria	71.5	28.5
Emilia-Romagna	71.2	28.8
Toscana	71.6	28.4
Umbria	65.9	34.1
Marche	60.9	39.1
Lazio	67.5	32.5
Abruzzo	59.6	40.4
Molise	58.6	41.4
Campania	59.2	40.8
Puglia	56.9	43.1
Basilicata	58.8	41.2
Calabria	57.6	42.4
Sicilia	59.7	40.3
Sardegna	69.5	30.5
Total	64.2	35.8

Simulations

Some preliminary simulations: Scenarios

- Scenario 1: $n = 1,000$; no selection bias

Some preliminary simulations: Scenarios

- Scenario 1: $n = 1,000$; no selection bias
- Scenario 2: $n = 2,000$; response rate $\approx 50\%$; selection bias due to differential nonresponse rate by sex, age, and educational level

Some preliminary simulations: Scenarios

- Scenario 1: $n = 1,000$; no selection bias
- Scenario 2: $n = 2,000$; response rate $\approx 50\%$; selection bias due to differential nonresponse rate by sex, age, and educational level
- 1,000 simulations for each scenario

Simulations results – Scenario 1

Mean n_d (n), θ_d (True), empirical standard error of $\hat{\theta}_d$ (Std), e.s.e. of $\tilde{\theta}_d$ (MrP)

Region	n	True	Std	MrP	MrP/Std %
Piemonte	93	33.2	4.9	2.6	52.9
Lombardia	95	37.9	4.9	2.9	59.9
Trentino-Alto Adige	51	45.3	6.8	5.4	80.4
Veneto	64	43.7	6.1	4.9	79.0
Friuli-Venezia Giulia	37	29.5	7.3	4.1	56.1
Liguria	38	26.7	7.3	3.6	49.8
Emilia-Romagna	59	25.7	5.8	3.5	59.5
Toscana	62	25.3	5.6	3.4	60.8
Umbria	30	30.3	8.4	3.9	45.9
Marche	39	38.8	7.8	3.3	42.5
Lazio	57	31.4	6.3	2.7	43.4
Abruzzo	43	38.8	7.8	2.9	37.3
Molise	29	39.6	9.4	2.8	29.5
Campania	60	43.0	6.5	2.9	45.1
Puglia	60	42.8	6.5	3.3	50.6
Basilicata	32	39.8	8.8	2.9	33.0
Calabria	46	40.0	7.1	3.1	43.1
Sicilia	61	40.9	6.2	2.7	44.5
Sardegna	44	31.2	6.9	2.6	37.3

Simulations results – Scenario 1

θ_d (True), bias of $\hat{\theta}_d$ (Std), bias of $\tilde{\theta}_d$ (MrP)

Region	True	Std	MrP
Piemonte	33.2	-0.2	0.5
Lombardia	37.9	0.1	-0.8
Trentino-Alto Adige	45.3	-0.1	-6.0
Veneto	43.7	0.1	-0.3
Friuli-Venezia Giulia	29.5	0.4	0.1
Liguria	26.7	-0.2	0.8
Emilia-Romagna	25.7	0.1	1.7
Toscana	25.3	0.1	1.9
Umbria	30.3	-0.1	1.5
Marche	38.8	0.1	-0.2
Lazio	31.4	0.2	1.5
Abruzzo	38.8	0.1	1.2
Molise	39.6	0.3	1.4
Campania	43.0	-0.4	-2.2
Puglia	42.8	-0.0	-0.5
Basilicata	39.8	0.1	1.0
Calabria	40.0	-0.1	1.4
Sicilia	40.9	0.0	-0.7
Sardegna	31.2	0.1	0.4

Simulations results – Scenario 1

θ_d (**True**), root mean square error of $\hat{\theta}_d$ (**Std**), rmse of $\tilde{\theta}_d$ (**MrP**)

Region	True	Std	MrP	MrP/Std %
Piemonte	33.2	4.9	2.7	54.0
Lombardia	37.9	4.9	3.0	62.3
Trentino-Alto Adige	45.3	6.8	8.1	120.0
Veneto	43.7	6.1	4.9	79.1
Friuli-Venezia Giulia	29.5	7.4	4.1	56.0
Liguria	26.7	7.3	3.7	51.0
Emilia-Romagna	25.7	5.8	3.9	66.0
Toscana	25.3	5.6	3.9	69.7
Umbria	30.3	8.4	4.2	49.4
Marche	38.8	7.8	3.3	42.6
Lazio	31.4	6.3	3.1	49.3
Abruzzo	38.8	7.8	3.2	40.5
Molise	39.6	9.4	3.1	33.0
Campania	43.0	6.5	3.7	56.4
Puglia	42.8	6.5	3.3	51.1
Basilicata	39.8	8.8	3.1	34.9
Calabria	40.0	7.1	3.4	47.3
Sicilia	40.9	6.2	2.8	45.8
Sardegna	31.2	6.9	2.6	37.8

Simulations results – Scenario 2

Mean n_d (n), θ_d (True), empirical standard error of $\hat{\theta}_d$ (Std), e.s.e. of $\tilde{\theta}_d$ (MrP)

Region	n	True	Std	MrP	MrP/Std %
Piemonte	89	33.2	5.0	2.7	53.5
Lombardia	93	37.9	5.2	3.0	58.1
Trentino-Alto Adige	48	45.3	7.4	5.8	77.5
Veneto	61	43.7	6.6	5.1	77.5
Friuli-Venezia Giulia	36	29.5	7.9	4.3	53.6
Liguria	39	26.7	7.2	3.6	50.3
Emilia-Romagna	60	25.7	6.0	3.5	58.0
Toscana	62	25.3	5.8	3.5	59.8
Umbria	31	30.3	8.6	4.1	47.7
Marche	38	38.8	7.9	3.5	44.6
Lazio	57	31.4	6.7	2.8	42.4
Abruzzo	42	38.8	8.1	2.9	35.6
Molise	28	39.6	9.4	2.9	30.4
Campania	58	43.0	6.7	3.0	45.1
Puglia	56	42.8	6.8	3.4	49.9
Basilicata	30	39.8	9.6	3.0	31.4
Calabria	44	40.0	7.7	3.2	41.5
Sicilia	57	40.9	6.6	3.0	44.8
Sardegna	40	31.2	7.5	2.8	36.9

Simulations results – Scenario 2

θ_d (True), bias of $\hat{\theta}_d$ (Std), bias of $\tilde{\theta}_d$ (MrP)

Region	True	Std	MrP
Piemonte	33.2	4.2	1.0
Lombardia	37.9	3.5	-0.5
Trentino-Alto Adige	45.3	2.8	-6.7
Veneto	43.7	3.4	-0.7
Friuli-Venezia Giulia	29.5	3.3	-0.1
Liguria	26.7	3.6	0.9
Emilia-Romagna	25.7	3.2	1.7
Toscana	25.3	3.6	1.9
Umbria	30.3	3.4	2.0
Marche	38.8	3.8	0.4
Lazio	31.4	3.8	1.7
Abruzzo	38.8	4.1	2.1
Molise	39.6	4.5	2.4
Campania	43.0	4.5	-1.2
Puglia	42.8	4.7	0.7
Basilicata	39.8	4.6	2.1
Calabria	40.0	4.5	2.5
Sicilia	40.9	4.1	0.2
Sardegna	31.2	5.2	0.9

Simulations results – Scenario 2

θ_d (`True`), root mean square error of $\hat{\theta}_d$ (`Std`), rmse of $\tilde{\theta}_d$ (`MrP`)

Region	True	Std	MrP	MrP/Std %
Piemonte	33.2	6.6	2.9	43.9
Lombardia	37.9	6.3	3.1	48.9
Trentino-Alto Adige	45.3	7.9	8.8	111.1
Veneto	43.7	7.4	5.1	69.6
Friuli-Venezia Giulia	29.5	8.6	4.3	49.5
Liguria	26.7	8.1	3.8	46.5
Emilia-Romagna	25.7	6.8	3.9	57.3
Toscana	25.3	6.8	4.0	58.3
Umbria	30.3	9.2	4.6	49.3
Marche	38.8	8.7	3.5	40.5
Lazio	31.4	7.7	3.3	42.7
Abruzzo	38.8	9.0	3.6	39.4
Molise	39.6	10.4	3.7	35.6
Campania	43.0	8.1	3.3	40.5
Puglia	42.8	8.3	3.5	42.1
Basilicata	39.8	10.7	3.7	34.4
Calabria	40.0	8.9	4.0	45.5
Sicilia	40.9	7.8	3.0	38.2
Sardegna	31.2	9.1	2.9	31.8

Some preliminary simulations: Results

- Scenario 1: compared to the standard ML estimator, the MrP estimator is more precise, even when n_d is relatively small

Some preliminary simulations: Results

- Scenario 1: compared to the standard ML estimator, the MRP estimator is more precise, even when n_d is relatively small
- Scenario 2: compared to the standard ML estimator, the MRP estimator is both more precise and less biased

Conclusion

Conclusion

- `mrp` is still at a very preliminary stage and it will take some time before it reaches a publishable form

Conclusion

- `mrp` is still at a very preliminary stage and it will take some time before it reaches a publishable form
- A second version of `mrp` will be submitted for presentation at the 2011 North American Stata Users Group Meeting

Conclusion

- `mrp` is still at a very preliminary stage and it will take some time before it reaches a publishable form
- A second version of `mrp` will be submitted for presentation at the 2011 North American Stata Users Group Meeting
- We also plan to write an article describing `mrp` and to submit it to *The Stata Journal*

References

References

- Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Gelman, A. and T.C. Little. 1997. Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* 23: 127–135.
- Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.
- Kastellec, J., Lax, J.R. and J.H. Phillips. 2010. Public opinion and Senate confirmation of Supreme Court nominees. *Journal of Politics* 72: 767–784.
- Lax, J.R. and J.H. Phillips. 2009a. How should we estimate public opinion in the States?. *American Journal of Political Science* 53: 107–121.
- Lax, J.R. and J.H. Phillips. 2009b. Gay rights in the States: Public opinion and policy responsiveness. *American Political Science Review* 103: 367–386.
- Park, D.K., Gelman, A. and J. Bafumi. 2004. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis* 12: 375–385.
- Park, D.K., Gelman, A. and J. Bafumi. 2006. State level opinions from national surveys: Poststratification using multilevel logistic regression. In *Public Opinion in State Politics*. Ed. J.E. Cohen. Stanford, CA: Stanford University Press, 209–228.