Multilevel Regression and Poststratification in Stata

Maurizio Pisati¹ Valeria Glorioso¹,²

maurizio.pisati@unimib.it v.glorioso@campus.unimib.it

¹Dept. of Sociology and Social Research
University of Milano-Bicocca (Italy)

²Dept. of Society, Human Development, and Health
Harvard School of Public Health

Stata Conference Chicago 2011
July 14-15
Outline

1 Introduction
   The problem
   The solution
Outline

1 Introduction
   The problem
   The solution

2 Stata command
Outline

1. Introduction
   - The problem
   - The solution

2. Stata command

3. Simulations
Outline

1. Introduction
   The problem
   The solution

2. Stata command

3. Simulations

4. Conclusion
Outline

1 Introduction
   The problem
   The solution

2 Stata command

3 Simulations

4 Conclusion

5 References
INTRODUCTION
A common research objective

- Sometimes social scientists are interested in determining whether, and to what extent, the distribution of a given target variable $Y$ varies across $K$ groups defined by the values of one or more covariates of interest.
A common research objective

- Sometimes social scientists are interested in determining whether, and to what extent, the distribution of a given target variable $Y$ varies across $K$ groups defined by the values of one or more covariates of interest.
- Let $G$ denote a discrete variable representing the $K$ groups under comparison. Without loss of generality, $G$ can represent either a single discrete covariate or the cross-classification of two or more discrete covariates.
A common research objective

• In symbols:

\[ G = \prod_{g=1}^{M_G} V_g \]

where \( \Pi \) is the Cartesian product operator; \( M_G \) denotes the total number of covariates forming \( G \); and \( V_g \) denotes the \( g^{th} \) covariate.
A common research objective

• In symbols:

\[ G = \prod_{g=1}^{M_G} V_g \]

where \( \Pi \) is the Cartesian product operator; \( M_G \) denotes the total number of covariates forming \( G \); and \( V_g \) denotes the \( g^{th} \) covariate

• We will refer to \( G \) as the *group variable*
A common research objective

- The (conditional) distribution of $Y$ within each category $k$ of $G$ can be described as follows:

$$Y_k \sim f(\theta_k, \phi_k) \quad \text{for } k = 1, \ldots, K$$

where $f(\cdot)$ denotes a generic probability distribution; $\theta_k$ denotes the expected value(s) of the distribution; and $\phi_k$ denotes one or more additional parameters of the distribution (e.g., its variance)
A common research objective

- For the sake of simplicity, let us focus on the expected value(s) of $Y$, so that our goal is to determine whether, and to what extent, the expected value(s) of $Y$ varies/vary across the $K$ categories of $G$. 
A common research objective

• For the sake of simplicity, let us focus on the expected value(s) of $Y$, so that our goal is to determine whether, and to what extent, the expected value(s) of $Y$ varies/vary across the $K$ categories of $G$

• In terms of regression analysis, this amounts to estimating the $K$ possible values of the regression function $E(Y|G = k)$, i.e., $E(Y|G = 1) \equiv \theta_1$, $E(Y|G = 2) \equiv \theta_2$, $\ldots$, $E(Y|G = K) \equiv \theta_K$
A common research objective

- For the sake of simplicity, let us focus on the expected value(s) of $Y$, so that our goal is to determine whether, and to what extent, the expected value(s) of $Y$ varies/vary across the $K$ categories of $G$.

- In terms of regression analysis, this amounts to estimating the $K$ possible values of the regression function $E(Y|G = k)$, i.e., $E(Y|G = 1) \equiv \theta_1$, $E(Y|G = 2) \equiv \theta_2$, \ldots, $E(Y|G = K) \equiv \theta_K$.

- Let us denote our estimand – i.e., our quantity of interest – by $\theta \equiv \{\theta_k : k = 1, \ldots, K\}$.
Estimating $\theta$

- How do we get accurate – i.e., precise and unbiased – estimates of $\theta$?
Estimating $\theta$

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

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

- The standard (maximum likelihood) estimator of each element $\theta_k$ of $\Theta$ is:

$$\hat{\theta}_k \equiv \mathbb{E}(Y|G = k) = \frac{\sum_{i=1}^{n_k} Y_i}{n_k}$$

where $n_k$ denotes the number of valid sample observations within category $k$ of variable $G$.
Estimating $\theta$

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


Estimating $\theta$

- When $n_k$ is small, $\hat{\theta}_k$ tends to be very unprecise, i.e., to generate highly variable estimates of $\theta_k$.
- The accuracy of $\hat{\theta}_k$ 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 $G$ is small and/or collected data are affected by selection bias, relatively accurate estimates of $\theta$ 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 $G$ is small and/or collected data are affected by selection bias, relatively accurate estimates of $\theta$ 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 $\theta$ – which we will denote by $\tilde{\theta}$ – can be described as a four-step procedure as follows:
The MRP estimator

- **First:** Identify one or more covariates that might possibly be responsible for selection bias. Without loss of generality, let $C$ denote a discrete variable representing the cross-classification of these covariates.

In symbols:

$$C = \prod_{c=1}^{M_C} V_c$$

where $\Pi$ is the Cartesian product operator; $M_C$ denotes the total number of covariates forming $C$; and $V_c$ denotes the $c^{th}$ covariate.

We will refer to $C$ as the *composition variable*
The MRP estimator

- **Second**: Define the new estimand $\gamma \equiv \{\gamma_{kl} : k = 1, \ldots, K; l = 1, \ldots, L\}$, where $\gamma_{kl} \equiv E(Y | G = k, C = l)$; $k$ indexes the $K$ categories of variable $G$ as above; and $l$ indexes the $L$ categories of variable $C$. 
• **Third:** Use a properly specified multilevel regression model to estimate $\gamma$
The MRP estimator

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

$$\tilde{\theta}_k = \sum_{l=1}^{L} \hat{\gamma}_{kl} w_{l|k}$$

where $w_{l|k} = N_{kl}/N_k$; $N_k$ denotes the number of members of the target population who belong in category $k$ of variable $G$; and $N_{kl}$ denotes the number of members of the target population who belong in category $k$ of variable $G$ and in category $l$ of variable $C$. 

Maurizio Pisati and Valeria Glorioso

Multilevel Regression and Poststratification 16/49
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 the composition variable $C$ 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 the composition variable $C$ is carefully defined, poststratification (step 4 above) helps to decrease bias.
- In sum, we expect MRP to be a relatively accurate estimator of $\theta$. 
The MRP estimator: Disadvantages

- We need to have population data – or, at least, a sufficiently accurate estimate of it – for the full $G \times C$ cross-classification; this might limit the definition of $C$. 

The MRP estimator: Disadvantages

- We need to have population data – or, at least, a sufficiently accurate estimate of it – for the full $G \times C$ cross-classification; this might limit the definition of $C$.
- To get good estimates of $\gamma$, the multilevel regression model must be specified very carefully – but this caveat applies to any kind of regression model.
STATA COMMAND
mrp – a Stata implementation of MRP

- mrp is a novel user-written Stata command that implements the MRP estimator outlined above
\textbf{mrp} – a Stata implementation of MRP

- \textbf{mrp} is a novel user-written Stata command that implements the MRP estimator outlined above.
- Basically, \textbf{mrp} requests the user to specify (a) the target variable $Y$; (b) the list of covariates forming the group variable $G$; (c) the list of covariates forming the composition variable $C$; (d) the multilevel regression command appropriate to the problem at hand (e.g., \texttt{xtmixed}); (e) the list of “fixed effects”; (f) the list of “random effects”; and (g) the name of a properly arranged dataset containing the population totals $N_{kl}$.
**mrp – a Stata implementation of MRP**

- **mrp** is a novel user-written Stata command that implements the MRP estimator outlined above.
- Basically, **mrp** requests the user to specify (a) the target variable $Y$; (b) the list of covariates forming the group variable $G$; (c) the list of covariates forming the composition variable $C$; (d) the multilevel regression command appropriate to the problem at hand (e.g., `xtmixed`); (e) the list of “fixed effects”; (f) the list of “random effects”; and (g) the name of a properly arranged dataset containing the population totals $N_{kl}$.
- The basic output of **mrp** is an estimate of the $K$ values of the regression function $E(Y|G = k)$, i.e., of the $K$ elements of $\theta$. 
Example (based on simulation)

- Our objective is to describe the extent to which the proportion of Italian adults who attend Catholic Mass regularly varies across Italian regions.
Our objective is to describe the extent to which the proportion of Italian adults who attend Catholic Mass regularly varies across Italian regions.

To this aim, a simple random sample of 2,000 units is drawn from the target population (Italian men and women aged 18+), and each sampled unit is contacted for interview.
Example (based on simulation)

- Our objective is to describe the extent to which the proportion of Italian adults who attend Catholic Mass regularly varies across Italian regions.
- To this aim, a simple random sample of 2,000 units is drawn from the target population (Italian men and women aged 18+), and each sampled unit is contacted for interview.
- Only 984 subjects accept to participate in the survey. The response rate turns out to be higher among women and positively correlated with age and educational level.
Example

- Since the number of valid observations within each region $k$ is generally small ($\min(n_k)=30$, $\max(n_k)=97$), the standard estimator of $\theta$ will be very unprecise.
Example

- Since the number of valid observations within each region $k$ is generally small ($\min(n_k)=30$, $\max(n_k)=97$), the standard estimator of $\theta$ will be very unprecise.
- Moreover, since sex, age, and educational level are associated with Catholic Mass attendance, the standard estimator of $\theta$ will likely be affected by selection bias.
Example

- Since the number of valid observations within each region $k$ is generally small ($\min(n_k)=30$, $\max(n_k)=97$), the standard estimator of $\theta$ will be very unprecise.

- Moreover, since sex, age, and educational level are associated with Catholic Mass attendance, the standard estimator of $\theta$ will likely be affected by selection bias.

- In an attempt to increase precision and decrease bias, we estimate $\theta$ using the new Stata command `mrp`.
Example: **mrp** specification

Target variable

```stata
mrp church, g(region relmar|region) c(sex age edu)
regcommand(xtmixed) binomial
fe(relmar) re(i.age i.edu i.sex i.region)
popref("PopRef.dta") npop(N)
percent
```
Example: \texttt{mrp} specification

List of covariates forming group variable $G$

\begin{verbatim}
  mrp church, g(region relmar|region) c(sex age edu) ///
  regcommand(xtmixed) binomial ///
  fe(relmar) re(i.age i.edu i.sex i.region) ///
  popref("PopRef.dta") npop(N) ///
  percent
\end{verbatim}
Example: **`mrp` specification**
List of covariates forming composition variable $C$

```
mrp church, g(region relmar|region) c(sex age edu) ///
  regcommand(xtmixed) binomial ///
  fe(relmar) re(i.age i.edu i.sex i.region) ///
  popref("PopRef.dta") npop(N) ///
  percent
```
Example: `mrp` specification
Multilevel regression command

```
  mrp church, g(region relmar|region) c(sex age edu) ///
  regcommand(xtmixed) binomial ///
  fe(relmar) re(i.age i.edu i.sex i.region) ///
  popref("PopRef.dta") npop(N) ///
  percent
```
Example: **mrp** specification

List of “fixed effects”

```
mrp church, g(region relmar|region) c(sex age edu) ///
regcommand(xtmixed) binomial ///
fe(relmar) re(i.age i.edu i.sex i.region) ///
popref("PopRef.dta") npop(N) ///
percent
```
Example: \texttt{mrp} specification

List of “random effects”

\begin{verbatim}
mrp church, g(region relmar|region) c(sex age edu) ///
regcommand(xtmixed) binomial ///
fe(relmar) re(i.age i.edu i.sex i.region) ///
popref("PopRef.dta") npop(N) ///
percent
\end{verbatim}
Example: **mrp specification**

Dataset and variable containing population totals $N_{kl}$

```stata
mrp church, g(region relmar|region) c(sex age edu) ///
regcommand(xtmixed) binomial ///
fe(relmar) re(i.age i.edu i.sex i.region) ///
popref("PopRef.dta") npop(N) ///
percent
```
Example: **mrp specification**
Scale option (converts proportions into percentages)

```
mrp church, g(region relmar|region) c(sex age edu) ///
regcommand(xtmixed) binomial ///
fe(relmar) re(i.age i.edu i.sex i.region) ///
popref("PopRef.dta") npop(N) ///
percent
```
Example: Results
Dot = True population value, S = Standard estimate, M = MrP estimate

\[ E(Y \mid G = k) \]
Simulations
Quantity of interest

- We used Monte Carlo simulation to evaluate the performance of three estimators of $\theta$ in a research setting analogous to the one illustrated in the example above.
Quantity of interest

- We used Monte Carlo simulation to evaluate the performance of three estimators of $\theta$ in a research setting analogous to the one illustrated in the example above.
- Our underlying research objective is to describe the extent to which the proportion of Italian adults who attend Catholic Mass regularly varies across 19 of the 20 regions into which Italy is subdivided (the 20th region, Valle d’Aosta, is excluded from the analysis because of its peculiarities).
Quantity of interest

- We used Monte Carlo simulation to evaluate the performance of three estimators of $\theta$ in a research setting analogous to the one illustrated in the example above.
- Our underlying research objective is to describe the extent to which the proportion of Italian adults who attend Catholic Mass regularly varies across 19 of the 20 regions into which Italy is subdivided (the 20th region, Valle d’Aosta, is excluded from the analysis because of its peculiarities).
- Thus, our quantity of interest $\theta$ corresponds to the $K = 19$ values of the regression function $E(Y|G = k)$, where the target variable $Y$ is a binary indicator of regular Catholic Mass attendance, and the group variable $G$ is the region of residence.
Procedure

• For each estimator, we followed a three-step procedure:
Procedure

- For each estimator, we followed a three-step procedure:
  1. First, we simulated 1,000 sample surveys, using as the sampling frame a large dataset \( N = 251,708 \) that mimics the socio-demographic structure of the full Italian adult population and contains complete information on the following individual characteristics: region of residence (\texttt{region}), sex (\texttt{sex}), age (\texttt{age}), educational level (\texttt{edu}), and Catholic Mass attendance (\texttt{church})
Procedure

For each estimator, we followed a three-step procedure:

1. First, we simulated 1,000 sample surveys, using as the sampling frame a large dataset \((N = 251,708)\) that mimics the socio-demographic structure of the full Italian adult population and contains complete information on the following individual characteristics: region of residence (\texttt{region}), sex (\texttt{sex}), age (\texttt{age}), educational level (\texttt{edu}), and Catholic Mass attendance (\texttt{church})

2. Second, we used the data collected in each simulated survey to estimate the quantity of interest, thus getting a simulated sampling distribution of \(\theta\) made of 1,000 estimates
Procedure

• For each estimator, we followed a three-step procedure:
  1. First, we simulated 1,000 sample surveys, using as the sampling frame a large dataset \( N = 251,708 \) that mimics the socio-demographic structure of the full Italian adult population and contains complete information on the following individual characteristics: region of residence (\texttt{region}), sex (\texttt{sex}), age (\texttt{age}), educational level (\texttt{edu}), and Catholic Mass attendance (\texttt{church})
  2. Second, we used the data collected in each simulated survey to estimate the quantity of interest, thus getting a simulated sampling distribution of \( \theta \) made of 1,000 estimates
  3. Finally, we evaluated the estimator in question by computing its bias, empirical standard error, and root mean square error
Survey specifications

- Sampling method: Simple random sampling
Survey specifications

• Sampling method: Simple random sampling
• Initial sample size: \( n = 2,000 \)
Survey specifications

- Sampling method: Simple random sampling
- Initial sample size: $n = 2,000$
- Response rate: Each sampled unit is selected into the final sample with a probability determined by his/her sex, age, and educational level. Such probabilities range from a minimum of 20% (poorly-educated men aged 18-44) to a maximum of 100% (highly-educated women aged 65+).
Survey specifications

- Sampling method: Simple random sampling
- Initial sample size: \( n = 2,000 \)
- Response rate: Each sampled unit is selected into the final sample with a probability determined by his/her sex, age, and educational level. Such probabilities range from a minimum of 20% (poorly-educated men aged 18-44) to a maximum of 100% (highly-educated women aged 65+)
- Final sample size: \( \text{mean}(n)=970, \min(n)=897, \max(n)=1,035 \)
Estimator 1
Standard (std)

• The standard estimator of each element $\theta_k$ of $\Theta$ is defined as follows:

$$\hat{\theta}_k = \frac{\sum_{i=1}^{n_k} \text{church}_i}{n_k}$$

where $\text{church}_i$ takes value 1 when subject $i$ attends Catholic Mass regularly, value 0 otherwise; and $n_k$ denotes the number of valid sample observations within region of residence $k$. 
• The MrP estimator of each element $\theta_k$ of $\Theta$ is defined as follows:

$$\tilde{\theta}_k = \sum_{l=1}^{L} \hat{\gamma}_{kl} w_{l|k}$$

where all symbols are defined as in slides 14-16 above
Estimator 2
Multilevel Regression with Poststratification (mrp)

- The MRP estimator of each element $\theta_k$ of $\Theta$ is defined as follows:

$$\tilde{\theta}_k = \sum_{l=1}^{L} \hat{\gamma}_{kl} w_{l|k}$$

where all symbols are defined as in slides 14-16 above.

- The estimation of parameters $\gamma_{kl}$ requires that the composition variable $C$ be previously defined.
Estimator 2
Multilevel Regression with Poststratification (mrp)

- The MrP estimator of each element $\theta_k$ of $\Theta$ is defined as follows:

$$\tilde{\theta}_k = \sum_{l=1}^{L} \hat{\gamma}_{kl} w_{l|k}$$

where all symbols are defined as in slides 14-16 above

- The estimation of parameters $\gamma_{kl}$ requires that the composition variable $C$ be previously defined

- In our case, we define $C$ as the cross-classification of three categorical covariates: sex (2 levels), age (4 levels), and edu (3 levels). Therefore, $L = 2 \times 4 \times 3 = 24$
• Given the definition of composition variable \( C \), the parameters \( \gamma_{kl} \) are estimated using the following multilevel regression model:

\[
\gamma_{kl} = \beta_0 + \alpha_{k}^{\text{region}} + \alpha_{r[l]}^{\text{sex}} + \alpha_{s[l]}^{\text{age}} + \alpha_{t[l]}^{\text{edu}}
\]
Estimator 2
Multilevel Regression with Poststratification (mrp)

where

\[
\alpha_k^{\text{region}} \sim N(\beta_{\text{relmar}} \cdot \text{relmar}, \sigma_{\text{region}}^2) \quad \text{for } k = 1, \ldots, 19
\]

\[
\alpha_r^{\text{sex}} \sim N(0, \sigma_{\text{sex}}^2) \quad \text{for } r = 1, \ldots, 2
\]

\[
\alpha_s^{\text{age}} \sim N(0, \sigma_{\text{age}}^2) \quad \text{for } s = 1, \ldots, 4
\]

\[
\alpha_t^{\text{edu}} \sim N(0, \sigma_{\text{edu}}^2) \quad \text{for } t = 1, \ldots, 3
\]

and \text{relmar} is a region-level variable that expresses the percentage of religious marriages in each region.
Estimator 3
Standard Regression with Poststratification (srp)

- The SrP estimator of each element \( \theta_k \) of \( \Theta \) is defined as follows:

\[
\bar{\theta}_k = \sum_{l=1}^{L} \hat{\gamma}_{kl} w_{l|k}
\]

where all symbols are defined as above
Estimator 3
Standard Regression with Poststratification (srp)

- The SRP estimator has the same general form as the MRP estimator, but in the SRP estimator the parameters $\gamma_{kl}$ are estimated using a standard logistic regression model as follows:

$$
\gamma_{kl} = \text{invlogit}(\beta_0 + \beta^{\text{relmar}} \cdot \text{relmar} + \beta^\text{region}_k \cdot \text{region}_k + \\
+ \beta^\text{sex}_r \cdot \text{sex}_r[l] + \beta^\text{age}_s \cdot \text{age}_s[l] + \beta^\text{edu}_t \cdot \text{edu}_t[l])
$$

where $\beta^\text{region}_1 = \beta^\text{region}_2 = \beta^\text{sex}_1 = \beta^\text{age}_1 = \beta^\text{edu}_1 = 0$
## Results: Bias

<table>
<thead>
<tr>
<th>Region</th>
<th>Method</th>
<th>$\theta_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piemonte</td>
<td>srp</td>
<td>33%</td>
</tr>
<tr>
<td>Lombardia</td>
<td>srp</td>
<td>39%</td>
</tr>
<tr>
<td>Trentino-Alto Adige</td>
<td>srp</td>
<td>50%</td>
</tr>
<tr>
<td>Veneto</td>
<td>srp</td>
<td>44%</td>
</tr>
<tr>
<td>Friuli-Venezia Giulia</td>
<td>mrp</td>
<td>29%</td>
</tr>
<tr>
<td>Liguria</td>
<td>srp</td>
<td>26%</td>
</tr>
<tr>
<td>Emilia-Romagna</td>
<td>srp</td>
<td>24%</td>
</tr>
<tr>
<td>Toscana</td>
<td>srp</td>
<td>25%</td>
</tr>
<tr>
<td>Umbria</td>
<td>mRp</td>
<td>30%</td>
</tr>
<tr>
<td>Marche</td>
<td>sRp</td>
<td>45%</td>
</tr>
<tr>
<td>Lazio</td>
<td>sRp</td>
<td>31%</td>
</tr>
<tr>
<td>Abruzzo</td>
<td>sRp</td>
<td>35%</td>
</tr>
<tr>
<td>Molise</td>
<td>sRp</td>
<td>40%</td>
</tr>
<tr>
<td>Campania</td>
<td>mRp</td>
<td>46%</td>
</tr>
<tr>
<td>Puglia</td>
<td>sRp</td>
<td>43%</td>
</tr>
<tr>
<td>Basilicata</td>
<td>sRp</td>
<td>36%</td>
</tr>
<tr>
<td>Calabria</td>
<td>sRp</td>
<td>40%</td>
</tr>
<tr>
<td>Sicilia</td>
<td>sRp</td>
<td>41%</td>
</tr>
<tr>
<td>Sardegna</td>
<td>sRp</td>
<td>31%</td>
</tr>
</tbody>
</table>

The diagram shows the results of bias for each region, with the method used (srp or mRp) and the corresponding $\theta_k$ value.
## Results: Empirical standard error

<table>
<thead>
<tr>
<th>Region</th>
<th>Method</th>
<th>Poststratification</th>
<th>Std</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piemonte</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>33%</td>
</tr>
<tr>
<td>Lombardia</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>39%</td>
</tr>
<tr>
<td>Trentino-Alto Adige</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>50%</td>
</tr>
<tr>
<td>Veneto</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>44%</td>
</tr>
<tr>
<td>Friuli-Venezia Giulia</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>29%</td>
</tr>
<tr>
<td>Liguria</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>26%</td>
</tr>
<tr>
<td>Emilia-Romagna</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>24%</td>
</tr>
<tr>
<td>Toscana</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>25%</td>
</tr>
<tr>
<td>Umbria</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>30%</td>
</tr>
<tr>
<td>Marche</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>45%</td>
</tr>
<tr>
<td>Lazio</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>31%</td>
</tr>
<tr>
<td>Abruzzo</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>35%</td>
</tr>
<tr>
<td>Molise</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>40%</td>
</tr>
<tr>
<td>Campania</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>46%</td>
</tr>
<tr>
<td>Puglia</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>43%</td>
</tr>
<tr>
<td>Basilicata</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>36%</td>
</tr>
<tr>
<td>Calabria</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>40%</td>
</tr>
<tr>
<td>Sicilia</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>41%</td>
</tr>
<tr>
<td>Sardegna</td>
<td>mrp</td>
<td>srp std</td>
<td>std</td>
<td>31%</td>
</tr>
</tbody>
</table>
## Results: Root mean square error

<table>
<thead>
<tr>
<th>Region</th>
<th>mrp</th>
<th>srp</th>
<th>std</th>
<th>θ_k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piemonte</td>
<td>33%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lombardia</td>
<td>39%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trentino-Alto Adige</td>
<td>50%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Veneto</td>
<td>44%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friuli-Venezia Giulia</td>
<td>29%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liguria</td>
<td>26%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emilia-Romagna</td>
<td>24%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toscana</td>
<td>25%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Umbria</td>
<td>30%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marche</td>
<td>45%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lazio</td>
<td>31%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abruzzo</td>
<td>35%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molise</td>
<td>40%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Campania</td>
<td>46%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puglia</td>
<td>43%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basilicata</td>
<td>36%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calabria</td>
<td>40%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sicilia</td>
<td>41%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sardegna</td>
<td>31%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

0 1 2 3 4 5 6 7 8 9 10 11


Maurizio Pisati and Valeria Glorioso
Results: Summary

- **Bias**: In absolute terms, the MrP estimator exhibits little bias – in most cases less than one percentage point, for an average true value of 36%. Comparatively, it exhibits significantly less bias than the standard estimator and slightly more bias than the SrP estimator.
Results: Summary

- **Bias**: In absolute terms, the MRP estimator exhibits little bias – in most cases less than one percentage point, for an average true value of 36%. Comparatively, it exhibits significantly less bias than the standard estimator and slightly more bias than the SRP estimator.

- **Precision**: The MRP estimator is significantly more precise (i.e., less variable) than both the standard estimator and the SRP estimator.
Results: Summary

- **Bias**: In absolute terms, the MRP estimator exhibits little bias – in most cases less than one percentage point, for an average true value of 36%. Comparatively, it exhibits significantly less bias than the standard estimator and slightly more bias than the SRP estimator.

- **Precision**: The MRP estimator is significantly more precise (i.e., less variable) than both the standard estimator and the SRP estimator.

- **Accuracy**: Combining bias and precision, we can conclude that the MRP estimator is 1 to 4 times more accurate than the standard estimator and 1 to 3 times more accurate than the SRP estimator.
CONCLUSION
• **mrp** is still at alpha stage and it will take a few months before it reaches a publishable form
• `mrp` is still at alpha stage and it will take a few months before it reaches a publishable form

• `mrp` is part of a larger project on the analysis of variation in Stata, and eventually it will be subsumed under a more general command for the analysis of association
Conclusion

- `mrp` is still at alpha stage and it will take a few months before it reaches a publishable form.
- `mrp` is part of a larger project on the analysis of variation in Stata, and eventually it will be subsumed under a more general command for the analysis of association.
- Part of the work presented here was carried out while Maurizio Pisati was a visiting scholar at the Institute for Quantitative Social Science at Harvard University, and Valeria Glorioso was a visiting student researcher at the Department of Society, Human Development, and Health of the Harvard School of Public Health.
## References
References