XVIII Italian Stata Conference
Florence, 9-10 May 2024

Non-Negative Least Squares in Stata

Giovanni Cerulli
IRCrES-CNR
Research Institute for Sustainable Economic Growth, National Research Council of Italy
E-mail: giovanni.cerulli@ircres.cnr.it

Non-Negative Least Squares (NNLS)

Given a matrix X of size $m \times n$ where each row represents a data point and each column represents a feature, and a vector y of size $m \times 1$ representing the observed dependent variable, we want to find a parameter vector β of size $n \times 1$ that minimizes the sum of squared errors subject to the constraint that all elements of β are nonnegative:

$$\min_{\beta} \|X\beta - y\|_2^2 \quad \text{subject to} \quad \beta \geq 0$$

The solution $\hat{\beta}$ to this minimization problem can be found using various algorithms tailored for nonnegative optimization.

Some NNLS applications

Spectral Analysis
In spectroscopy, signals are often nonnegative, such as spectra representing the intensity of light at different wavelengths.

Chemometrics
NNLS is widely used in chemometrics for tasks like multivariate calibration in analytical chemistry.

Image Processing
In image processing, NNLS can be applied for tasks like image reconstruction from incomplete or noisy data.

Biomedical Signal Processing
NNLS has applications in processing various biomedical signals, such as electroencephalography (EEG) and positron emission tomography (PET) data.

Text Mining and Natural Language Processing
In topic modeling and document clustering, NNLS can be used to decompose document-term matrices into nonnegative components, facilitating the discovery of latent topics or themes in text collections.

Environmental Monitoring
NNLS can be applied in environmental monitoring for source apportionment and pollution source identification. By decomposing pollution data into nonnegative components, it becomes easier to identify and quantify the contributions of different pollution sources.

Financial Modeling
In finance, NNLS can be utilized for portfolio optimization, where the weights assigned to different assets need to be nonnegative. It can also be applied in risk management and asset pricing models.

NNLS in Stata: the **nnls** command

help nnls

Title

nnls —— Non-negative least squares in Stata

Syntax

nnls *depvar indepvars* [*if*] [*in*], [**graph graph_save(***graphname***) standardize**]

options	Description
graph	generates the importance index bar graph
standardize	compute NNLS on z-standardized variables (zero mean and unit variance)

Example of `nnls` application

```
Example: Non-negative least squares applied to the Boston dataset
    Load initial dataset from ancillary file
        use boston, clear
    Set the outcome
        global y "medv"
    Set the features
        global X "crim zn indus age lstat black"
    Run "nnls" using unstandardized variables
        nnls $y $X , graph graph_save("my_graph")
    Generate predictions
        predict PRED_ustd
    Clear all the Stata environment
        clear all
    Load again initial dataset from ancillary file
        use boston, clear
    Run "nnls" using standardized variables
        nnls $y $X , graph graph_save("my_graph") standardize
    Generate predictions
        predict PRED_std
```

NNLS in Stata - results

Results using **variables in levels** (original variables)

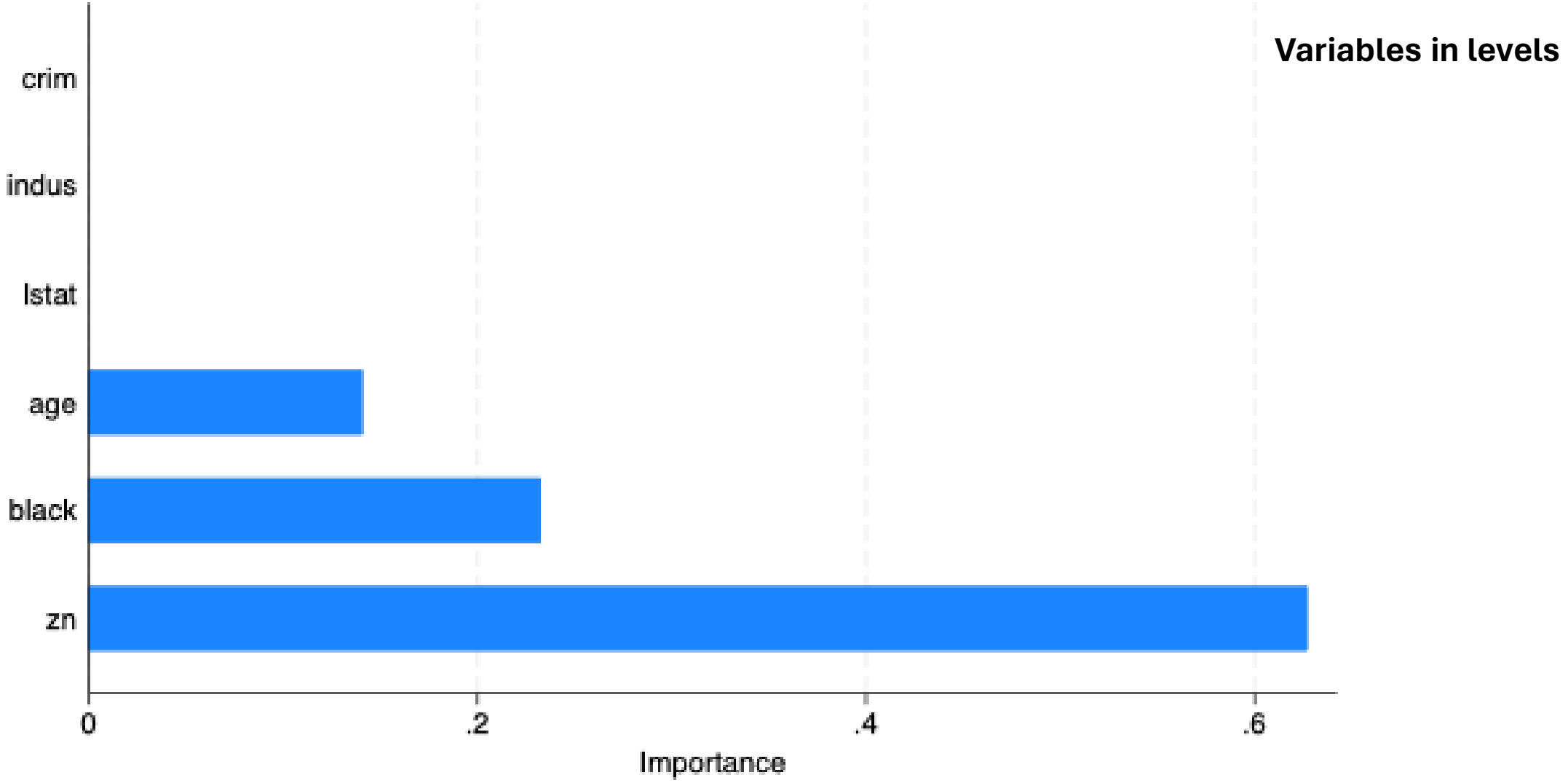
WEIGHTS: not summing up to 1

	Weights
crim	0
zn	.1386488
indus	0
age	.0310916
lstat	0
black	.0513815

WEIGHTS: summing up to 1

	Unit_We~s
crim	0
zn	.6270242
indus	0
age	.1406083
lstat	0
black	.2323675

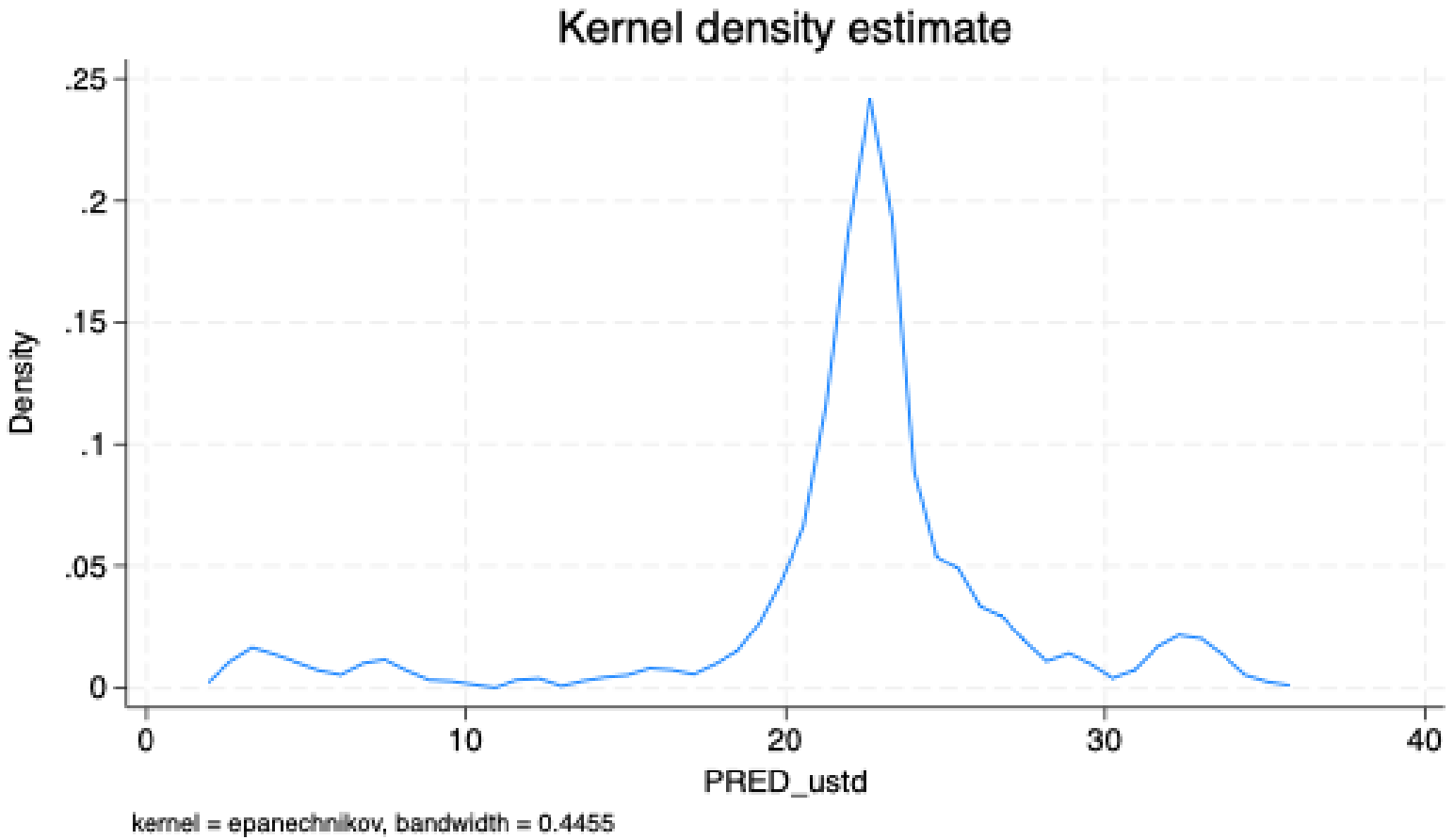
Feature importance indicators



Making NNLS predictions

```
. predict PRED_ustd
. kdensity PRED_ustd
```

Variables in levels



NNLS in Stata - results

Results using **standardized variables** (zero mean, unit variance)

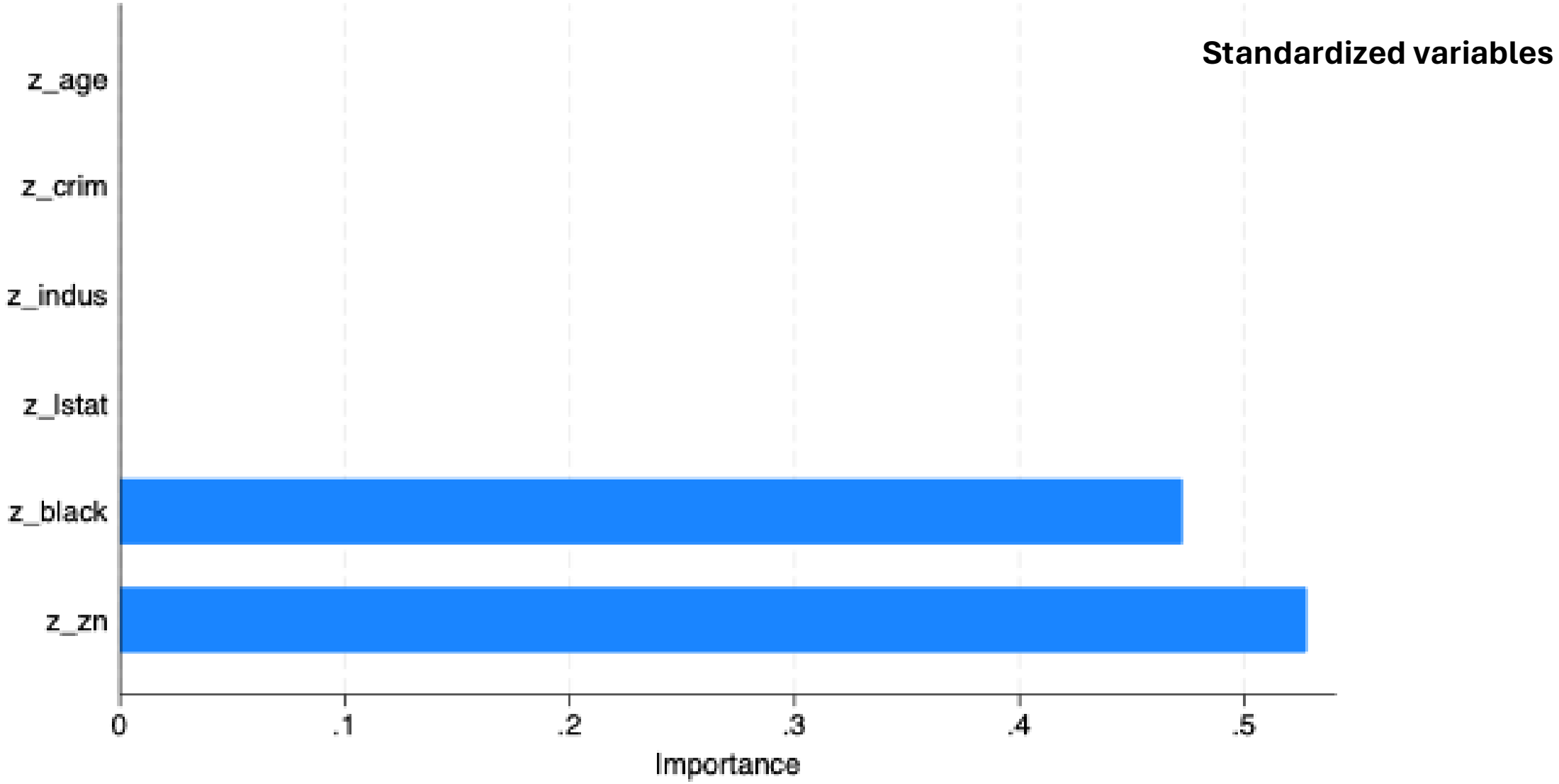
WEIGHTS: not summing up to 1

	Weights
z_crim	0
z_zn	.3115131
z_indus	0
z_age	0
z_lstat	0
z_black	.2787839

WEIGHTS: summing up to 1

	Unit_We~s
z_crim	0
z_zn	.5277226
z_indus	0
z_age	0
z_lstat	0
z_black	.4722774

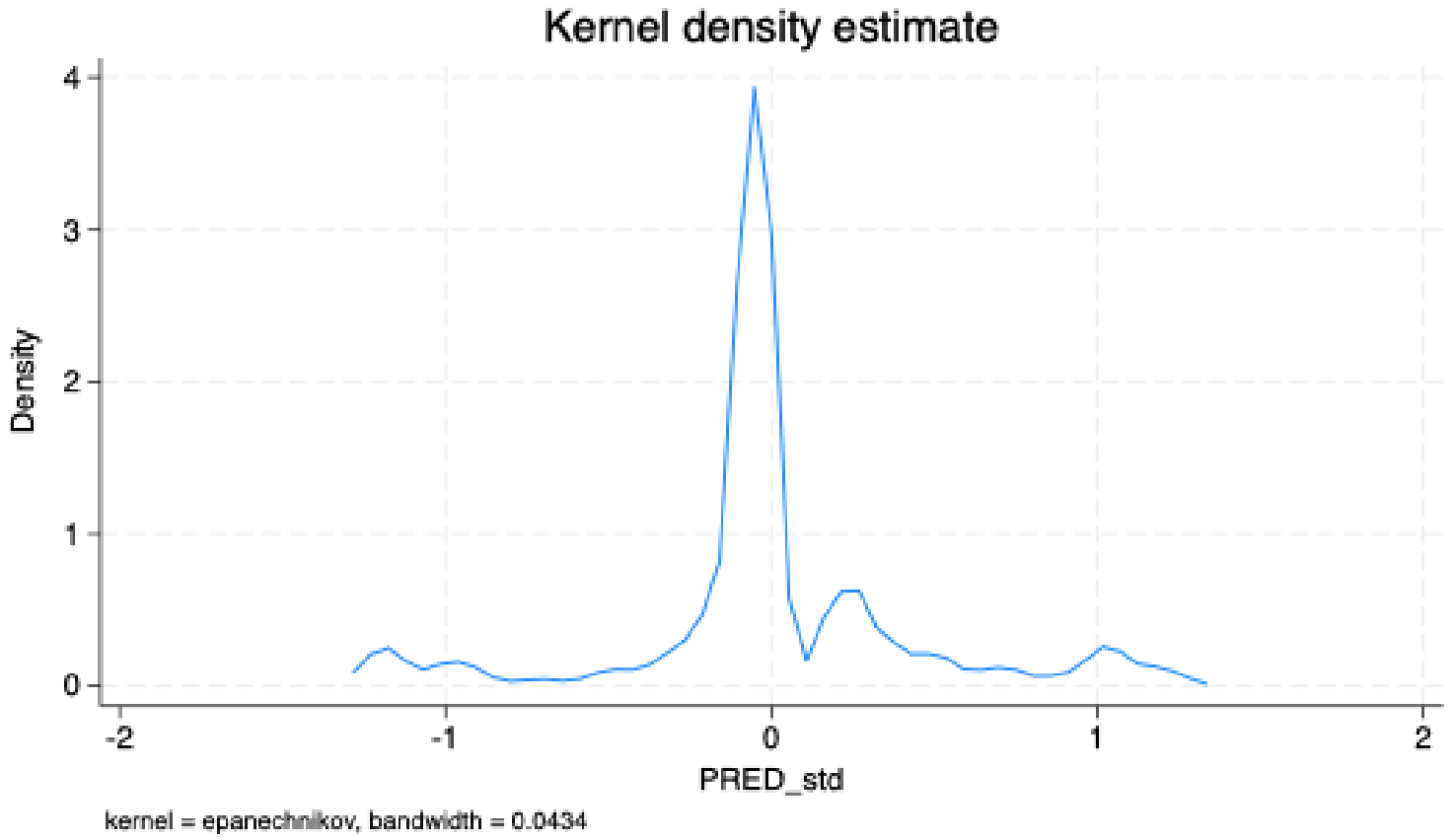
Feature importance indicators



Making NNLS predictions

```
. predict PRED_std
. kdensity PRED_std
```

Standardized variables



Remark: Python installation needed

```
Remark 1: In order to execute this program, it is necessary to have both Stata 16 or newer versions and Python
    installed, starting from version 2.7 onwards. Detailed instructions for installing Python on your machine
    can be found at: https://www.  rg/downloads. We highly recommend using the Anaconda distribution for
    Python, which can be installed from: https://docs.anaconda.com/free/anaconda/install/index.html#.
    Additionally, prior to running the command, it is essential to ensure that the "SciPy" package and its
    related dependencies, as well as the Stata Function Interface (sfi) APIs, are installed.
```

Conclusions

References

Bro, R. and De Jong, S. 1997. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics*, 11, 393-401.

Cerulli, G. 2023. *Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata*, Springer.

Acknowledgment

The development of this software was supported by FOSSR (Fostering Open Science in Social Science Research), a project funded by the European Union - NextGenerationEU under the NPRR Grant agreement n. MURIR0000008.