wqsreg - A Stata command for Weighted Quantile Sum regression

Marta Ponzano[1], Stefano Renzetti[2], Andrea Bellavia[3]

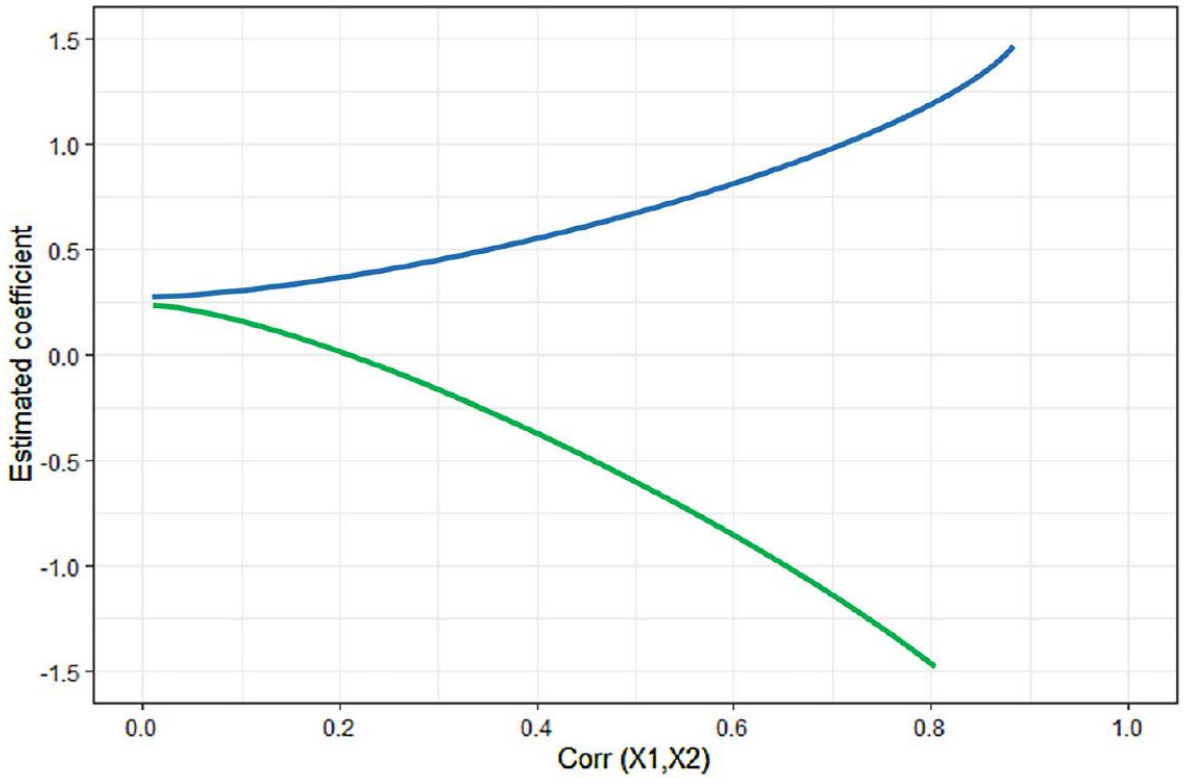
[1] Department of Health Sciences, University of Genoa, Genoa, Italy

[2] Department of Medicine and Surgery, University of Parma, Parma, Italy

[3] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA

Correlated predictors

- The presence of highly correlated predictors is a well-known issue in numerous domains
- Including highly correlated covariates in the same regression model increases the risk of multicollinearity:
  - ↓ accuracy
  - ↑ standard errors



The reversal paradox.

Correlated predictors – environmental epidemiology

- This issue is particularly common in environmental epidemiology, when studying the health effects of a combined group of multiple factors, known as environmental mixtures:

1. Mixture Bad Actors (specific role of each factor)
2. Cumulative Mixture Effect (overall)

1. 2.

Statistical methods

- To address this complexity, some methodologies have been considered (e.g., penalized regression) and *ad hoc* statistical methods have been developed and then used also in different fields outside of environmental epidemiology:

  - Mixture Indexing approaches

    Weighted Quantile Sum (WQS) Regression, Quantile G-computation, etc.
  - Flexible Approaches for complex settings

    Bayesian Kernel Machine Regression (BKMR), Machine Learning approaches, etc.

WQS Regression

- WQS regression is a supervised statistical approach where exposure aggregation via indexing is conducted in two steps:

  1. Creating a summary index by weighting each mixture component
  2. Including the index in a regression model

1. 2.

1. Creating the summary index - weights estimation

- Split the dataset into training and validation (40/60)
- Score all exposure factors into categories based on quantiles (quartiles)
- Generate *b* bootstrap samples of the training dataset (100)
- For each sample, estimate parameters (weights and βs):

$$f(y) = \beta_0 + \beta_1 \left(\sum_{i=1}^{c} w_i q_i \right) + \boldsymbol{\beta \cdot C'}$$

$0 \leq w_i \leq 1$; $\sum_{i=1}^{c} w_i = 1$

Assumption of unidirectionality (pre-specified direction): only average results over bootstrap samples with either positive or negative estimates

- Derive $\overline{w}_i$ for each component based on estimates from bootstrap samples
- Derive the final estimate of the WQS as weighted sum of the components

$$WQS = \sum_{i=1}^{c} \bar{w}_i q_i$$

2. Including the index in a regression model

- Fit a regression model in the validation set to estimate the βs

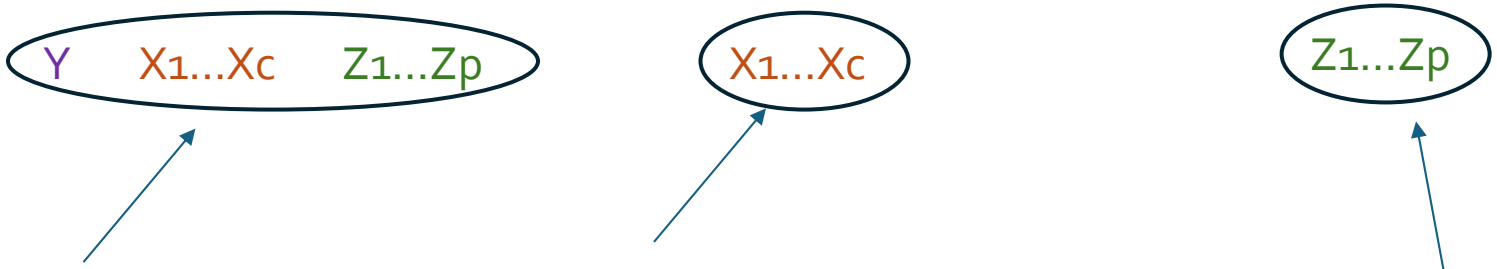
$$f(y) = \beta_0 + \beta_1 \cdot WQS + \boldsymbol{\beta} \cdot \mathbf{C}'$$

WQS Regression in Stata

A Stata command has not been implemented yet. To address this gap, we have developed a new command – **wqsreg**

- ***wqsreg*** fits WQS regression for continuous outcomes, while allowing for the several flexible components of this framework. It returns the regression estimates, the weights, and the WQS index for each subject.

wqsreg -Syntax



Syntax:

```
wqsreg depvar indepvars, mixture(varlist)  boot(integer) ///
[, validation(integer 0) q(integer 4) b1_neg(integer 0) cvar(varlist) seed(integer
0) conv_maxiter(integer 2000) conv_vtol(real 0.000000001) technique(string)
saveWQSindex(integer 0) saveWeights(integer 0) datasetWQSindexName(string)
datasetWeightsName(string) figureName(string) id(string)]
```

wqsreg -Syntax

Number of bootstrap samples (>0).
boot=1: no bootstrapping

Syntax:

```
wqsreg depvar indepvars, mixture(varlist)  boot(integer) ///
[, validation(integer 0) q(integer 4) b1_neg(integer 0) cvar(varlist)
seed(integer 0) conv_maxiter(integer 2000) conv_vtol(real 0.000000001)
technique(string) saveWQSindex(integer 0) saveWeights(integer 0)
datasetWQSindexName(string) datasetWeightsName(string) figureName(string)
id(string)]
```

wqsreg -Syntax

Syntax:

```
wqsreg depvar indepvars, mixture(varlist)  boot(integer) ///
[, validation(integer 0) q(integer 4) b1_neg(integer 0) cvar(varlist)
seed(integer 0) conv_maxiter(integer 2000) conv_vtol(real 0.000000001)
technique(string) saveWQSindex(integer 0) saveWeights(integer 0)
datasetWQSindexName(string) datasetWeightsName(string) figureName(string)
id(string)]
```

% Validation set [0; 100)

Determining the quantile

Unidirectionality:
0: Positive
1: Negative

wqsreg -Syntax

Syntax:

```
wqsreg depvar indepvars, mixture(varlist)  boot(integer) ///
[, validation(integer 0) q(integer 4) b1_neg(integer 0) cvar(varlist)
seed(integer 0) conv_maxiter(integer 2000) conv_vtol(real 0.000000001)
technique(string) saveWQSindex(integer 0) saveWeights(integer 0)
datasetWQSindexName(string) datasetWeightsName(string) figureName(string)
id(string)]
```

Seed

WQS index

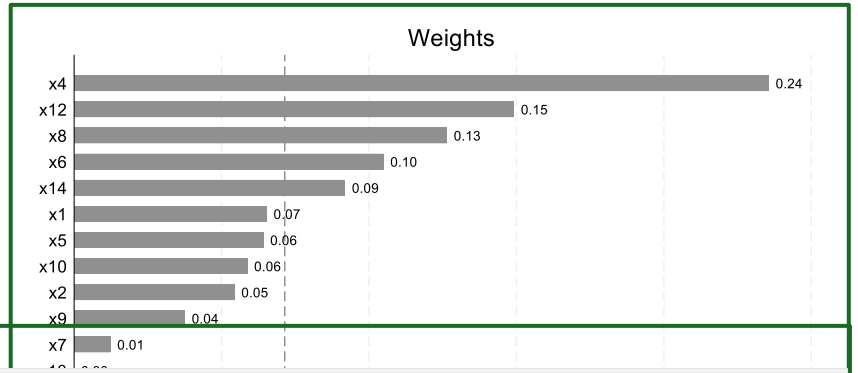
	Obs	Validation	WQS_index
1	1	1	1.448031
2	2	0	2.986457
3	3	1	2.332614
4	4	1	2.724882
5	5	1	3.034041
6	6	1	2.870383
7	7	0	2.259726

wqsreg -Syntax

Syntax:

```
wqsreg depvar indepvars, mixture(varlist)  boot(integer) ///
[, validation(integer 0) q(integer 4) b1_neg(integer 0) cvar(varlist)
seed(integer 0) conv_maxiter(integer 2000) conv_vtol(real 0.000000001)
technique(string) saveWQSindex(integer 0) saveWeights(integer 0)
datasetWQSindexName(string) datasetWeightsName(string) figureName(string)
id(string)]
```

Weights



	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
1	.0652756	.0543328	8.08e-12	.2355599	.0641369	.1048592	.0122648	.1263925	.0375168	.05868	8.97e-12	.1491447	9.05e-12	.0918369

wqsreg -Syntax

Syntax:

```
wqsreg depvar indepvars, mixture(varlist)  boot(integer) ///
[, validation(integer 0) q(integer 4) b1_neg(integer 0) cvar(varlist)
seed(integer 0) conv_maxiter(integer 2000) conv_vtol(real 0.000000001)
technique(string) saveWQSindex(integer 0) saveWeights(integer 0)
datasetWQSindexName(string) datasetWeightsName(string) figureName(string)
id(string)]
```

Technique:
Broyden–Fletcher–Goldfarb–Shanno (bfgs)
Modified Newton–Raphson (nr)

Maximum number of iterations to be performed before optimization

Tolerance

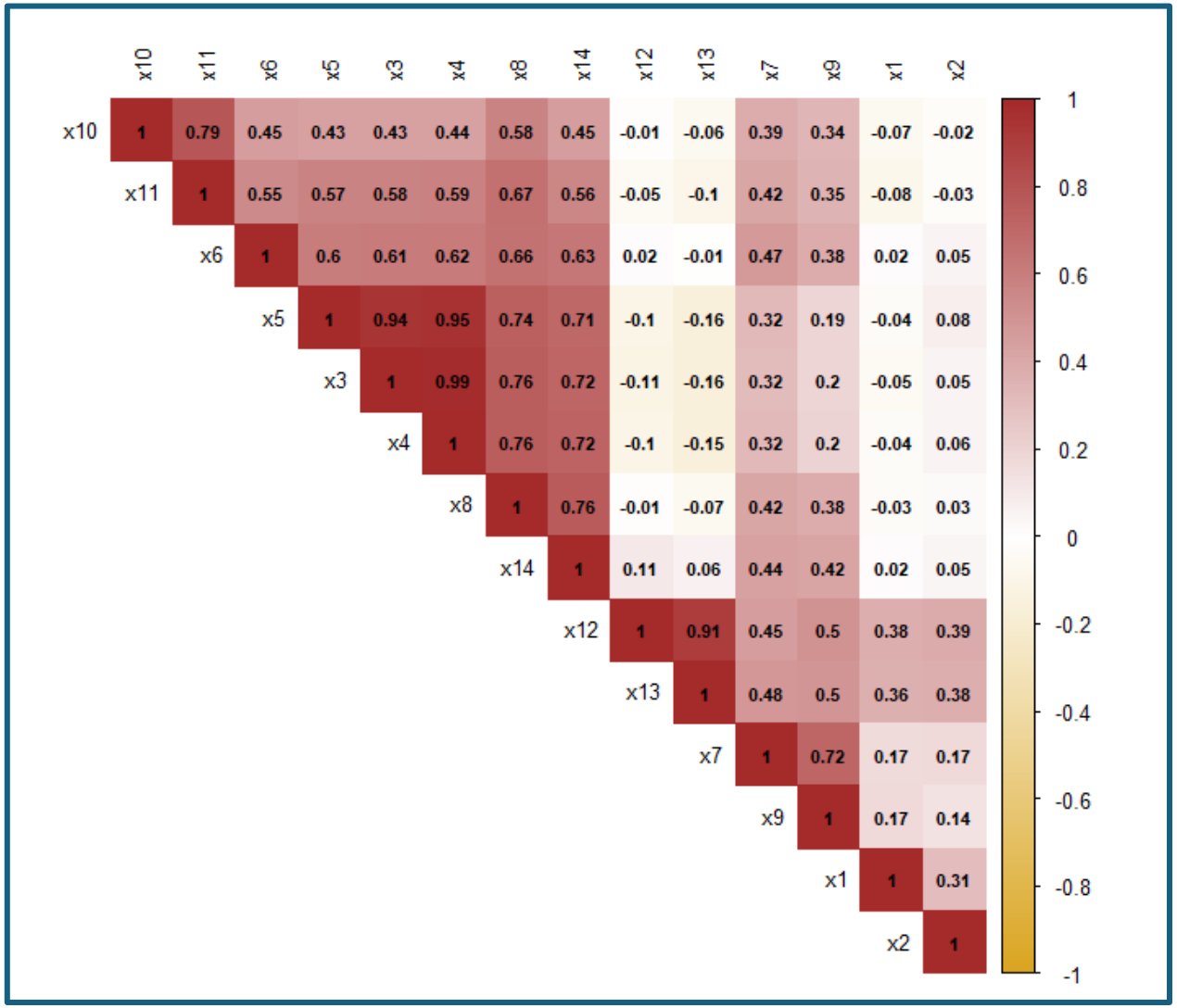
Examples - Simulated data set

- The National Institute of Environmental Health Sciences held a workshop in 2015 to address the need to develop novel statistical approaches for multi-pollutant epidemiology studies.

- Participants were asked to analyze simulated data sets made available six months before.

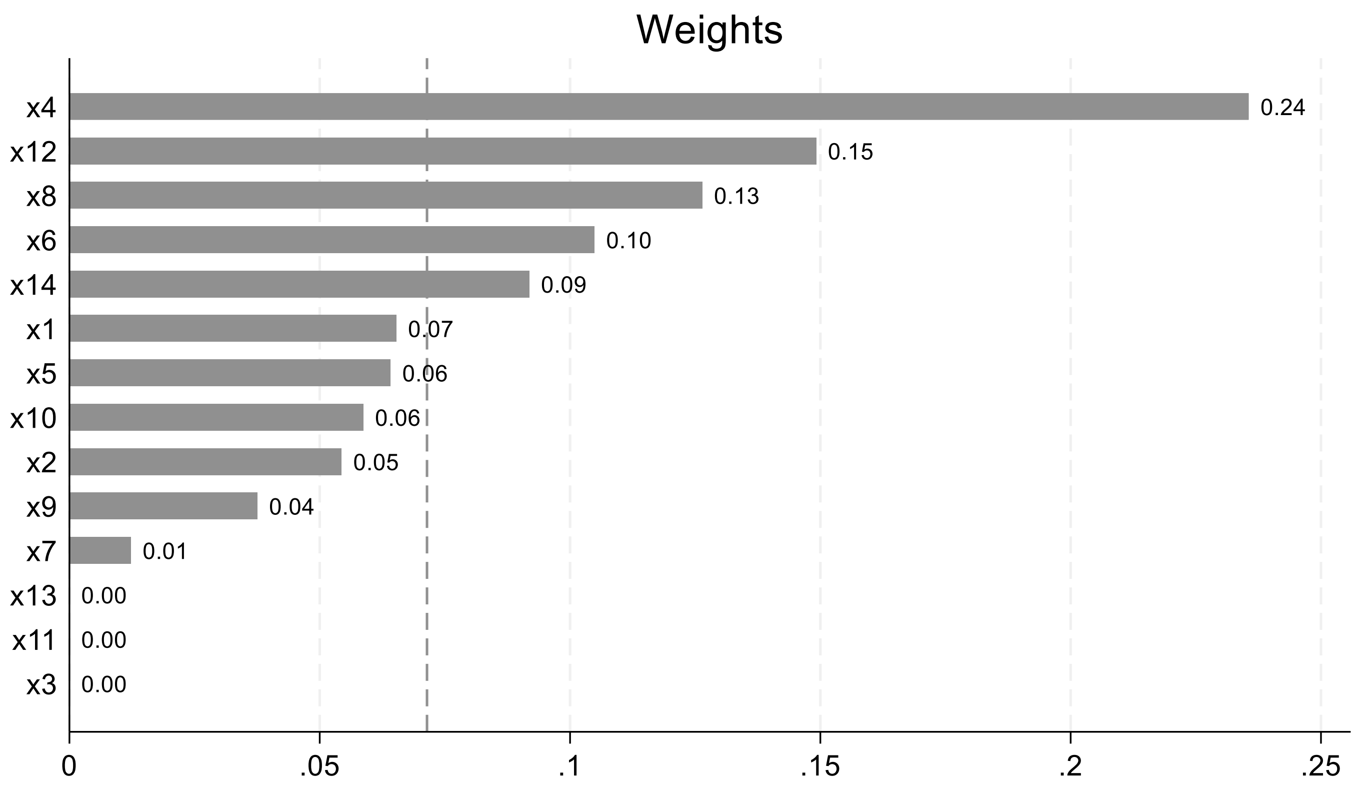
  - Simulated data set #2 (n = 500) designed to represent a cross-sectional study of 14 exposure variables:
    - Clusters of highly correlated covariates
    - $E[Y] = 3 + 0.05\mathbf{X4} + 0.1\mathbf{X6} + 0.1\mathbf{X11} + 0.5\mathbf{X12} + 0.1\mathbf{X14} + 0.01Z1 + 0.003Z2$ for $Z3=0$
    - $E[Y] = 2.68 + 0.01\mathbf{X1} + 0.05\mathbf{X4} + 0.1\mathbf{X11} + 0.1\mathbf{X14} + 0.01Z1 + 0.003Z2$ for $Z3=1$

Examples - Simulated data set



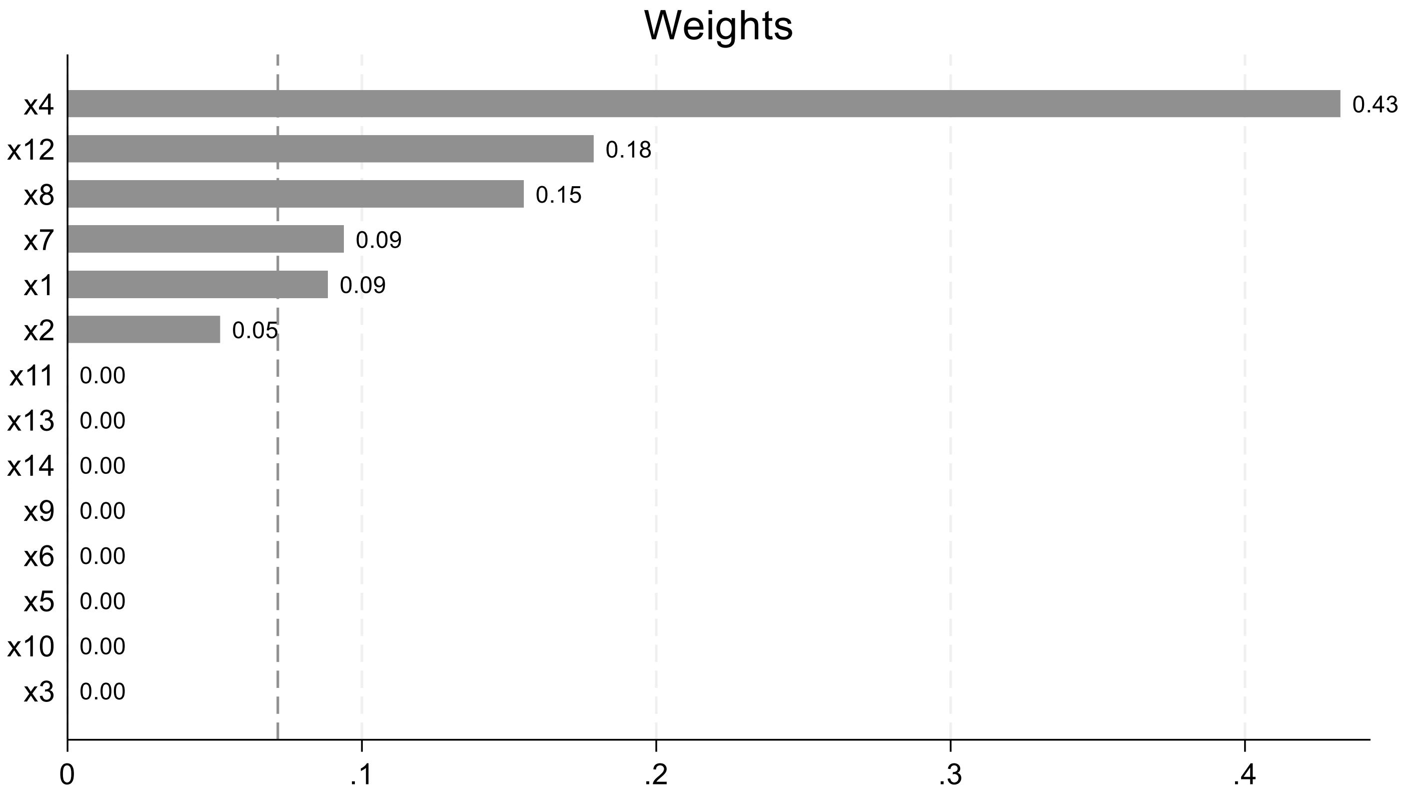
Example - Mixture of 14 Components – no bootstrap and no training/validation split

```
. wqsreg y x*, mixture(x*) boot(1)

-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*.
N observations used - Total: 500
N observations used - Validation: 500
WQS index Coef: .51137906
Std. Err.: .0349305
t-value: 14.639901
p-value: 1.273e-40
95% CI: [.44274975, .58000838]
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*.
```



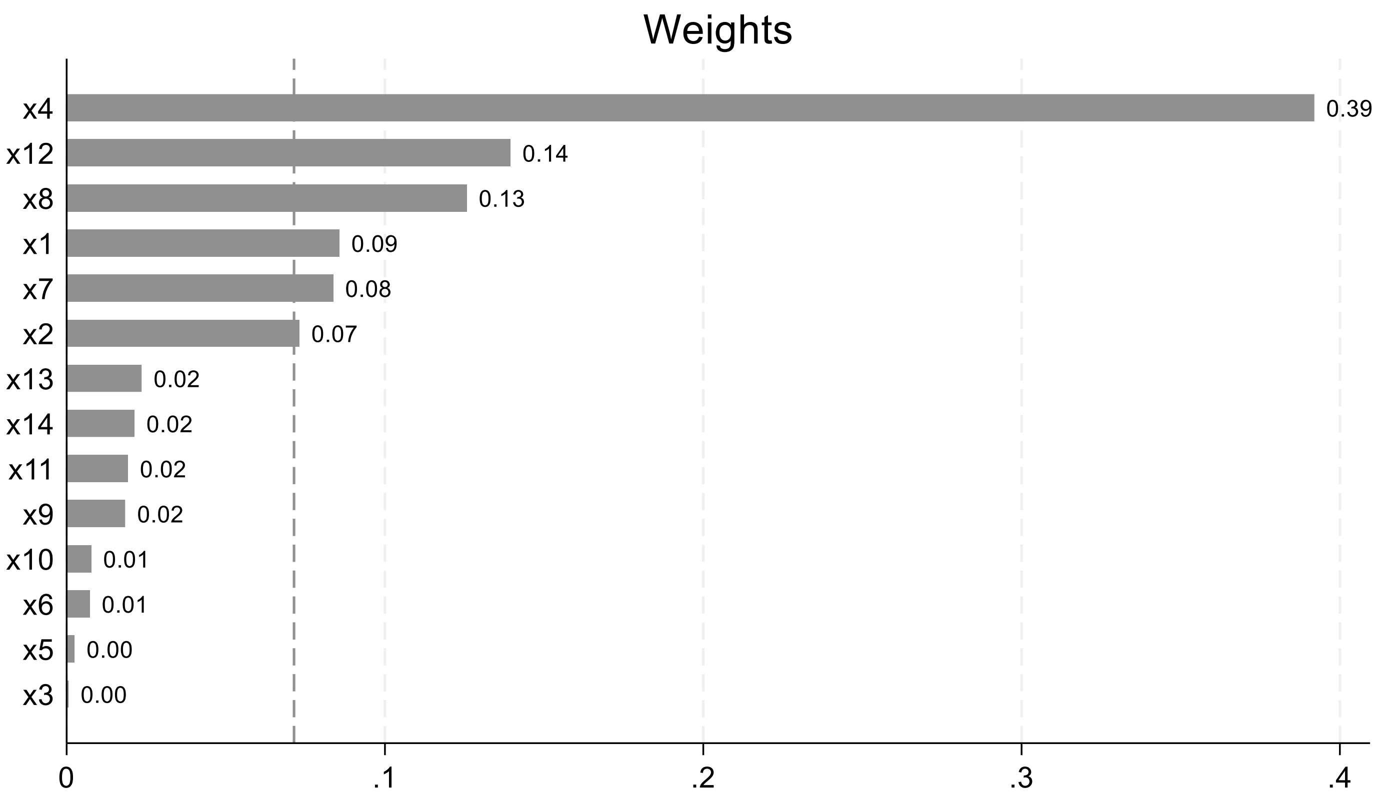
Example - Mixture of 14 Components – no bootstrap, 60% validation

```
. wqsreg y x*, mixture(x*) boot(1)  validation(60)

-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*.
N observations used - Total: 500
N observations used - Validation: 290
WQS index Coef: .45213493
Std. Err.: .04585987
t-value: 9.8590544
p-value: 6.027e-20
95% CI: [.36187192, .54239793]
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*.
```



Example - Mixture of 14 Components – 100 bootstrap samples, 60% validation

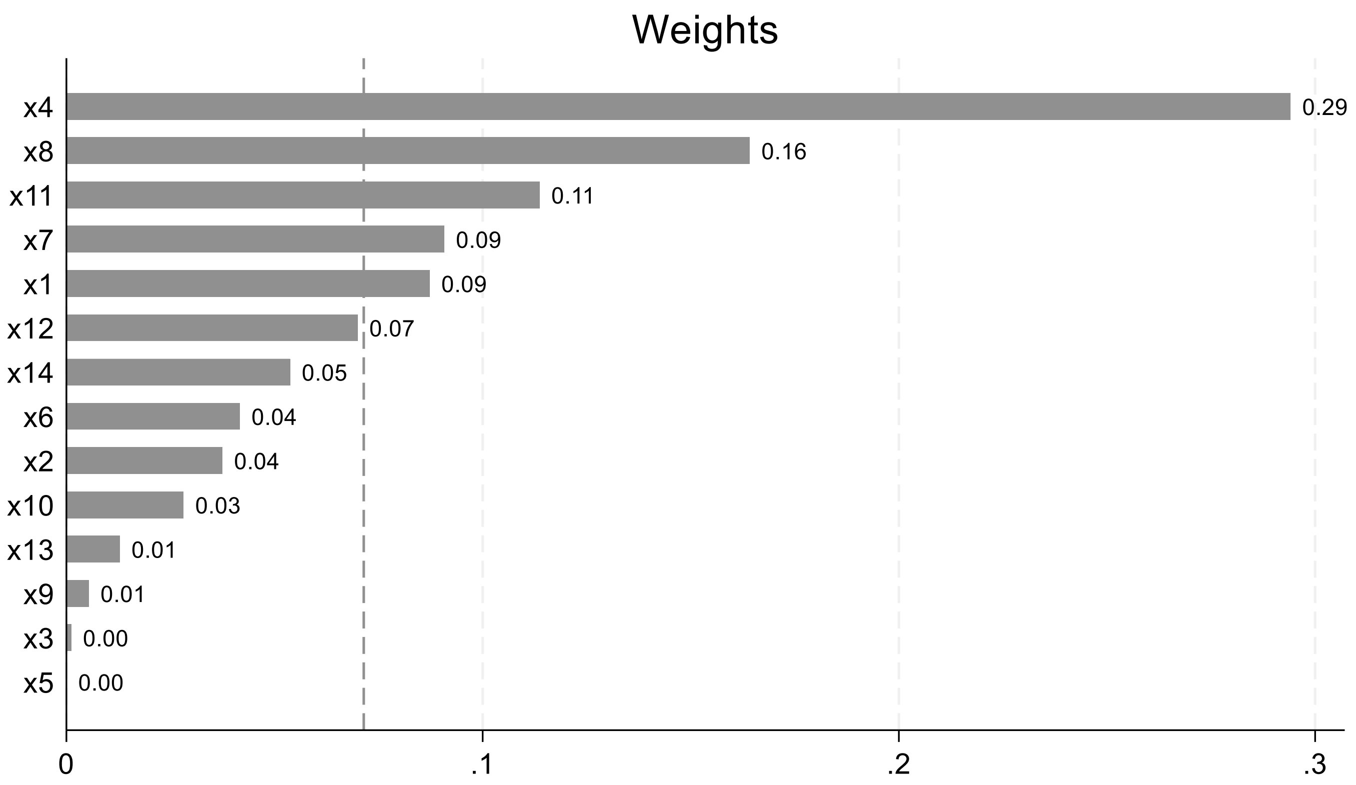
```
. wqsreg y x*, mixture(x*) boot(100)  validation(60)

-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-
N observations used - Total: 500
N observations used - Validation: 290
WQS index Coef: .47321036
Std. Err.: .04704726
t-value: 10.058193
p-value: 1.359e-20
95% CI: [.3806103, .56581043]
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-
```



Example - Mixture of 14 Components – Z3-adjusted

```
. wqsreg y x* z3, mixture(x*) boot(100)  validation(60) cvar(z3)
```

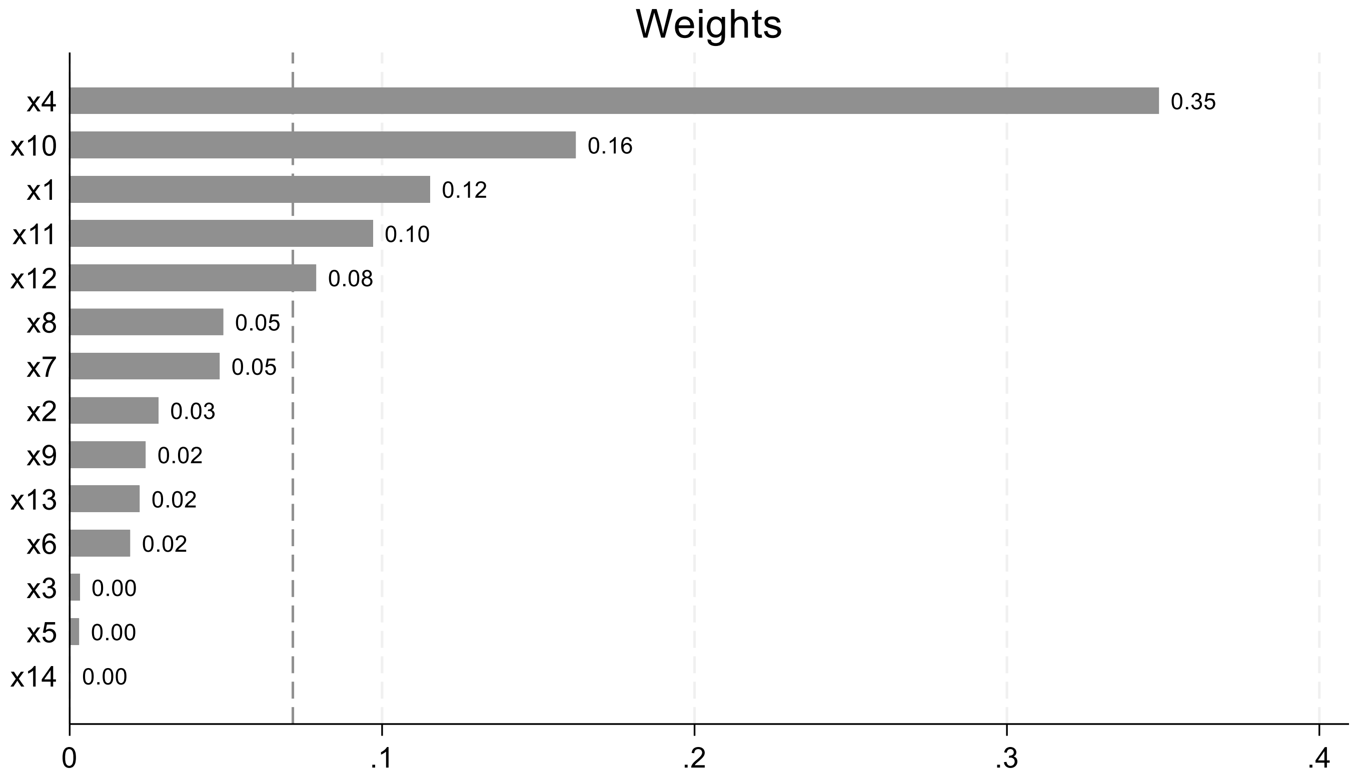
```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-
N observations used - Total: 500
N observations used - Validation: 290
WQS index Coef: .40892724
Std. Err.: .0379083
t-value: 10.787275
p-value: 5.336e-23
95% CI: [.3343137, .48354078]
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-
```



Example - Mixture of 14 Components – Z3-adjusted seed(1994)

```
. wqsreg y x* z3, mixture(x*) boot(100)  validation(60) cvar(z3) seed(1994)
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
N observations used - Total: 500
N observations used - Validation: 314
WQS index Coef: .3869652
Std. Err.: .0400741
t-value: 9.6562427
p-value: 1.822e-19
95% CI: [.30811456, .46581584]
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
```



Possible errors generated by ***wqsreg*** include:

```
. wqsreg y x*, mixture(x*) boot(1)  b1_neg(1)
Error: There are no negative b1
```

```
. wqsreg  y x*, mixture(x*)  b1_neg(2)  boot(1)
Error, b1_neg can be 0 (positive b1, the default) or 1 (negative b1)

. wqsreg  y x*, mixture(x*)  b1_neg(0)  boot(1) validation(-5)
Error, Validation must be numeric in [0; 100)

. wqsreg  y x* z1, mixture(x*)  b1_neg(0)  boot(1)
Error, please check the number of mixture components and of confounders

. wqsreg  y x* z1, mixture(x*)  b1_neg(0)  boot(1) cvar(z1 z2)
Error, please check the number of mixture components and of confounders

. wqsreg  y x* z1, mixture(x*)  b1_neg(0)  boot(0)
Error, please insert a postive number of bootstrap samples. Note that boot=1 means no bootstrapping

. wqsreg  y x* z1, mixture(x*)  b1_neg(0)  boot(0)    datasetWQSindexName("Hello")
Error, datasetWQSindexName is allowed only when saveWQSindex=1
```

Next steps:

- Extensions:
  - Generalization to binary and poisson
  - Repeated holdout
  - Random subset

Discussion

- The importance of appropriately exploring complex multidimensional exposures, such as environmental mixtures, is increasingly recognized.
- Although further steps can be implemented, ***wqsreg*** is the first command to apply WQS regression in Stata.

Selected references

- Bellavia A., Statistical Methods for Environmental Mixtures, Springer, 2025
- Carrico, C., Gennings, C., Wheeler, D. C., & Factor-Litvak, P. (2015). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. Journal of agricultural, biological, and environmental statistics, 20(1), 100-120.
- Czarnota, J., Gennings, C., & Wheeler, D. C. (2015). Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk. Cancer informatics, 14, CIN-S17295.
- Renzetti, S., Curtin, P., Just, A. C., Bello, G., Gennings, C., Renzetti, M. S., & Rsolnp, I. (2021). Package 'gWQS'.

ponzano.marta@gmail.com