MethodsX 6 (2019) 2855–2860

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

Repeated holdout validation for weighted quantile sum regression

Eva M. Tanner[a,*], Carl-Gustaf Bornehag[a,b], Chris Gennings[a]

[a] Icahn School of Medicine at Mount Sinai, New York, NY, United States
[b] Karlstad University, Karlstad, Sweden

A B S T R A C T

Weighted Quantile Sum (WQS) regression is a method commonly used in environmental epidemiology to assess the impact of chemical mixtures in relation to a health outcome of interest. Data are partitioned into a single training and test set to reduce sample-specific chemical weights. However, in typical epidemiology sample sizes, this may produce unstable chemical weights and WQS index estimates, and investigators may resort to training and testing on the same data. To solve this problem, we propose repeated holdout validation whereby data are randomly partitioned 100 times, producing a distribution of validated results. Taking the mean as the final estimate, confidence estimates may also be calculated for inference. Further, this method helps characterize the variability in chemical weights, aiding in the identification of chemicals of concern. This is important since it may direct future research into specific chemicals.

Using data from 718 mother-child pairs in the Swedish Environmental Longitudinal, Mother and Child, Asthma and Allergy (SELMA) study, we assessed the association between prenatal exposure to 26 endocrine disrupting chemicals and child Intelligence Quotient (IQ). Results using a single partition were unstable, varying by random seed. The WQS index estimate was significant when all data was used (e.g. no partition) ($\beta = -2.2$ CI = -3.43, -0.98), but attenuated and nonsignificant using repeated holdout validation ($\beta = -0.82$ CI = -2.11, 0.45). When implementing WQS in epidemiologic studies with limited sample sizes, repeated holdout validation is a viable alternative to using a single, or no partitioning. Repeated holdout can both stabilize results and help characterize the uncertainty in identifying chemicals of concern, while maintaining some of the the rigor of holdout validation.

- Repeated holdout validation improves the stability of WQS estimates in finite study samples
- Uncertainty in identifying toxic chemicals of concern is acknowledged and characterized

A R T I C L E I N F O
Method name: Repeated holdout validation for weighted quantile sum regression
Keywords: Environmental epidemiology, Chemical mixtures, Cross-validation, Bootstrap, Uncertainty plot, Chemical of concern
Article history: Received 18 September 2019; Accepted 6 November 2019; Available online 22 November 2019

DOI of original article: http://dx.doi.org/10.1016/j.envint.2019.105185
* Corresponding author.
E-mail address: eva.tanner@mssm.edu (E.M. Tanner).

https://doi.org/10.1016/j.mex.2019.11.008

Specification Table

Subject Area:	Environmental Science
More specific subject area:	Environmental Epidemiology
Method name:	Repeated Holdout Validation for Weighted Quantile Sum Regression
Name and reference of original method:	Weighted Quantile Sum Regression Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. 2015. Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. J Agric Biol Environ Stat 20:100–120; doi:10.1007/s13253-014-0180-3.
Resource availability:	gWQS R Package (https://cran.r-project.org/web/packages/gWQS/index.html) Repeated_Holdout_WQS code (http://doi.org/10.5281/zenodo.2658697)

Method details

Weighted Quantile Sum (WQS) regression is an approach used in environmental epidemiology to evaluate associations between potentially highly correlated co-exposures and a health outcome [1]. Exposure values are quantiled and combined into a unidirectional weighted index, thereby reducing dimensionality and avoiding multi-collinearity. WQS provides a single overall effect estimate of the mixture that is easier to interpret than many other mixtures methods, and individual chemicals are ranked by their overall contribution to the index, indicating relative importance. In simulations, WQS demonstrated improved accuracy over traditional regression and shrinkage methods [1]. More recently, extensions to WQS have enabled wider applications, including interaction and stratification [2,3], high-dimensional data [4], and the distributed lag modeling framework for serial exposure measurements [5].

Equation 1 shows the WQS regression formula [1]. For $j = 1$ to c components of exposure, q_{ji} is the quantile of component j for the ith individual. The weight w_j is estimated for each of the j components, where weights take on values between 0 and 1 and sum to 1. WQS regression analysis is conducted in multiple steps. First, weights are estimated using a nonlinear modeling algorithm where the regression coefficients and weights are estimated simultaneously. An ensemble step is added for stabilization – e.g., weights are estimated across bootstrapped samples and the final weights are determined by their average [6]. The overall effect of the mixture (WQS index) is estimated by β_1, with weights constrained to a single direction, and is linked to the mean outcome μ_i using a generalized linear model, along with the intercept β_0, matrix of covariates z_i' and their corresponding coefficients φ.

$$g(\mu_i) = \beta_0 + \beta_1 \left(\sum_{j=1}^{c} w_j q_{ji} \right) + z_i' \varphi \tag{1}$$

The index can be estimated in both positive and negative directions with separate constrained analyses in the nonlinear estimation step. The constraint of focusing the inference in a single direction (combined with the constraints that the weights sum to 1) has the advantage of improving the ill-conditioning of the estimation due to complex correlation patterns in the quantiled components. The ensemble step provides the advantage of stabilizing the weights while accommodating variability in their estimates. Ideally, WQS uses a training set for the model fitting in ensemble steps, and conducts a hypothesis test on the WQS index in a holdout, or validation set. Finally, when two indices are estimated using constraints in the positive direction and one in the negative direction, they may be combined in a final model to evaluate their joint relationship with the mean response.

Validation techniques are important tools used in predictive modeling and machine learning to evaluate the replicability of results [7]. Even when prediction, variable selection, or model selection is not the goal, validation can help assess the generalizability and stability of findings [7,8]. Most previous WQS regression applications partitioned data into a single training and test set to avoid sample-specific chemical weights and WQS index estimates (Fig. 1), which may partly reflect random

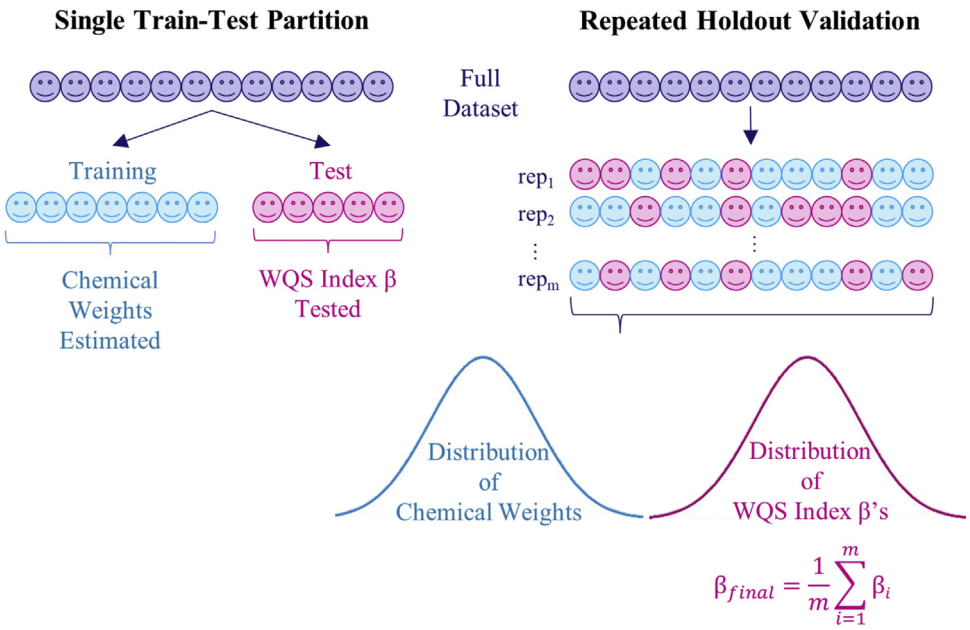

Fig. 1. Comparison of Standard versus Novel Partitioning Schemes for WQS.
Conventional WQS regression partitions a full dataset into a single training and test set to estimate chemical weights and test the association between the WQS index and outcome (left). Repeated holdout validation randomly partitions data m times and takes the average WQS index estimate (right).

noise [1]. However, in finite study samples this reduces statistical power and may lead to unrepresentative partitions and unstable estimates [9]. While stratified random partitioning can produce balanced partitions based on a categorical variable of interest, this procedure is less practical when analyzing multiple continuous chemical exposure variables in WQS regression. Because of this instability, investigators may forgo partitioning, training and testing on the same full dataset. However, we show that this may produce optimistic results.

To overcome this problem, we implemented repeated holdout validation which combines cross-validation and bootstrap resampling [9]. Specifically, we randomly partitioned (with replacement) the dataset 100 times and repeated WQS regression on each set to simulate a distribution of validated results from the underlying population (Fig. 1). Within each repetition, we still included the bootstrap step endorsed by Carrico et al. [1] to ensure weights within a single training partition were stable with improved sensitivity and specificity. With 100 bootstraps per repetition and 100 repetitions, weights were estimated 10,000 times. Therefore, a drawback is that this procedure is more computational intensive, taking 100 times longer to run compared to typical WQS implementations. The distribution of 100 validated results approximated the normal distribution in our analysis of 718 subjects. However, a larger number of repetitions would provide better normal approximations (e.g. $\geq$ 1000 repetitions as is typical for bootstrapping) [10]. Note that the training-test split percentages are somewhat arbitrary; we used 40%/60% training-testing splits as suggested by Carrico et al. [1] to provide additional power to the test set for testing the significance of the beta parameter, as compared to a 50%/50% split. We conducted analysis in R (R [11]) using the gWQS package [12] and provide additional code for conducting repeated holdout validation and compiling results in GitHub repository [13].

From the simulated distributions, we took the mean as the final estimate for the chemical weights and WQS index β coefficient. For coefficient inference, we calculated the 95 % confidence intervals (CI) based on the standard deviation (SD) of the simulated sampling distribution since this corresponds to the standard error (SE) calculated for a single sample [10]. Note that the SE is much smaller than the SD

in the simulated distribution and would give unreasonably narrow CIs. Although unconventional, we did this to facilitate comparison with results from training and testing on the full dataset which are reported using symmetric CIs.

Our example data comes from a study of prenatal exposure to 26 endocrine disrupting chemicals in relation to child Intelligence Quotient (IQ) among mother-child pairs from the Swedish Environmental Longitudinal, Mother and Child, Asthma and Allergy (SELMA) study [14]. Chemicals included triclosan, bisphenols A, F, and S (BPA, BPF, BPS), monoethyl, monobutyl, monobenzyl, di-2-ethylhexyl, diisononyl, monohydroxyisodecyl, and monocarboxyisononyl phthalates (MPE, MBP, MBzP, DEHP, DINP, MHiDP, MCiNP), 2-4-methyl-7-oxyooctyl-oxycarbonyl-cyclohexane carboxylic acid (MOiNCH), diphenylphosphate (DPHP), 3,5,6-trichloro-2-pyridinol (TCP), 3-phenoxybenzoic acid (PBA), 2-hydroxyphenanthrene (2OHPH), perfluorooctanoic acid (PFOA), perfluorooctane sulfonate (PFOS), perfluorononanoic acid (PFNA), perfluorodecanoic acid (PFDA), perfluoroundecanoic acid (PFUnDA), perfluorohexane sulfonic acid (PFHxS), hexachlorobenzene (HCB), trans-nonachlor (Nonachlor), dichlorodiphenyltrichloroethane and its metabolite dichlorodiphenyldichloroethylene summed (DDT), and 10 summed polychlorinated biphenyls (PCB). We set the chemical of concern threshold to a weight of 3.8 %, a value consistent with equal weighting (100 %/26 chemicals).

Compared to running WQS on the full dataset without validation, repeated holdout results were attenuated towards the null and nonsignificant (Table 1). This does not indicate that results obtained without validation are incorrect, but that they may only apply to that specific study sample, and may not generalize. The machine learning literature calls this resubstitution error, and is known to give overly-optimistic results [9]. Inference from sampling distributions typically uses percentile-based estimates and CIs (e.g. 2.5^{th}, 50^{th}, 97.5^{th} centiles). We observed similar results using either of the estimate and CI derivations (Table 1).

Another advantage of repeated holdout validation is that it allows the investigator to characterize weight uncertainty, aiding in the identification of toxic chemicals of concern. We created a weight uncertainty plot which efficiently displays all distributional information (Fig. 2). The bars correspond to the right axis and show the number of repetitions a chemical weight surpassed the chemical of concern threshold of 3.8 % out of the 100 repeated holdouts. All other plot information corresponds to the left axis, indicating actual weights (expressed as percentages) with the threshold value clearly marked. Boxplots display the 25^{th}, 50^{th}, and 75^{th} centiles, with whiskers indicating the 10^{th} and 90^{th} centiles. Diamonds display mean weights. Individual data points display the weights from each repetition.

Extreme individual weights exemplify why single partitions may lead to incorrect conclusions regarding a particular chemical. For example, DPHP had the second highest mean weight (10 %) in the WQS index, but seven of 100 repetitions were below the chemical of concern threshold, demonstrating that it may have been misclassified if only one partition was analyzed. Conversely, the mean weight for Triclosan (3 %) was below the chemical of concern threshold, but 26 % of repetitions had weights above the threshold. This may be due to random error or an unmeasured confounder related to Triclosan and IQ. This demonstrates one aspect of why a chemical may be related to neurodevelopmental outcomes in some studies, but not others. The simulated distribution allows the investigator to evaluate how replicable results may be if the study were repeated using a new sample from the same underlying population, or another population with similar demographics and chemical exposure patterns.

There are alternatives to repeated holdout for WQS, but they may only be suitable for specific research questions. K-fold cross-validation partitions data into 5–10 folds, allowing the WQS index

Table 1
WQS Index β Coefficients and CIs by Validation Technique & Estimation Type.

Validation Technique	Estimation Type	β Coefficient	Lower Limit	Upper Limit
None: Train/Test Full Dataset	Mean & SE-based 95 % CI	−2.20	−3.43	−0.98
Repeated Holdout	Mean & SD-based 95 % CI	−0.83	−2.11	0.45
Repeated Holdout	Median, 2.5^{th} & 97.5^{th} percentiles	−0.86	−1.99	0.43

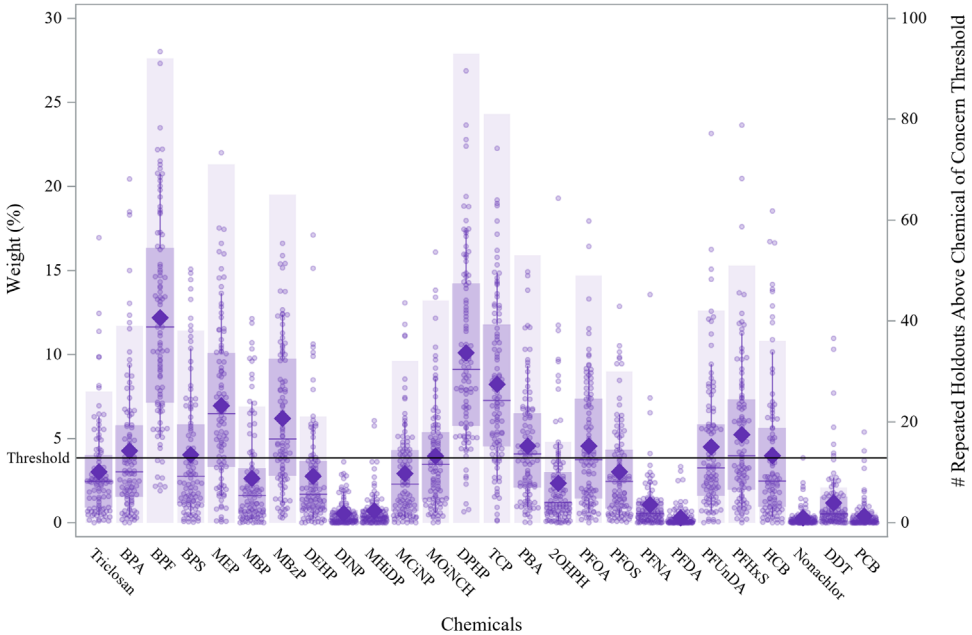

Fig. 2. Chemicals of Concern Identification & Uncertainty for 26 Endocrine Disrupting Chemicals in Relation to IQ. Bars correspond to right axis and indicate the number of times a chemical exceeded the concern threshold in 100 repeated holdouts. Data points, boxplots, and diamonds correspond to left axis. Data points indicate weights for each of the 100 holdouts. Box plots show 25th, 50th, and 75th percentiles, and whiskers show 10th and 90th percentiles of weights for the 100 holdouts. Closed diamonds show mean weights for the 100 holdouts. For comparison, open diamonds show the mean weight of the full sample analysis. Threshold = 3.8 %

estimate to be averaged across the partitions. In contrast to repeated holdout, it guarantees that each subject is rotated through training and test sets. However, k-fold validation is more appropriate when the goal is predictive accuracy, whereas the primary focus of WQS regression is chemical weight sensitivity and specificity. In high dimensional mixtures settings, WQS with random subsetting (WQSRS) may be used [4]. This method iteratively selects random subsets of exposures and combines results across multiple ensemble steps. Simulations showed that WQSRS performed well with over 400 predictor variables.

Conclusion

Training and testing on the same dataset is consistent with most epidemiology studies, but this methodology has limitations that are seldom acknowledged. Specifically, we may simply be fitting to random noise despite our best efforts to control for the many biases inherent in observational studies. Compared to training and testing on the same dataset, using a validation hold-out set to test for significance of the WQS index helps achieve a higher level of rigor, with results that may be more generalizable and repeatable. Using a single partition for training and validation is appropriate when the sample size is large enough to produce stable results regardless of random seeds. In smaller samples, repeated holdout validation can produce more stable WQS index estimates, and help characterize the uncertainty in the selection of chemicals of concern. Repeated holdout validation is a useful extension to WQS regression, allowing an investigator to retain some of the rigor of holdout testing in epidemiologic-relevant sample size.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the EDC-MixRisk (634880) European Union's Horizon 2020 Research and Innovation Programme and the National Institute of Environmental Health Sciences Powering Research Through Innovative Methods for Mixtures in Epidemiology (PRIME) Program (R01ES028811-01).

References

[1] C. Carrico, C. Gennings, D.C. Wheeler, P. Factor-Litvak, Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting, J. Agric. Biol. Environ. Stat. 20 (2015) 100–120, doi:http://dx.doi.org/10.1007/s13253-014-0180-3.
[2] M.J. Lee, M.H. Rahbar, M. Samms-Vaughan, J. Bressler, M.A. Bach, M. Hessabi, M.L. Grove, S. Shakespeare-Pellington, C. Coore Desai, J.A. Reece, K.A. Loveland, E. Boerwinkle, A generalized weighted quantile sum approach for analyzing correlated data in the presence of interactions, Biom. J. 61 (2019) 934–954, doi:http://dx.doi.org/10.1002/bimj.201800259.
[3] S. Renzetti, C. Gennings, P. Curtin, gWQS: An R Package for Linear and Generalized Weighted Quantile Sum (WQS) Regression [WWW Document] URL, (2019) . https://cran.r-project.org/web/packages/gWQS/vignettes/gwqs-vignette.pdf.
[4] P. Curtin, J. Kellogg, N. Cech, C. Gennings, A random subset implementation of weighted quantile sum (WQS RS) regression for analysis of high-dimensional mixtures, Commun. Stat. Simul. Comput. (2019) 1–16, doi:http://dx.doi.org/10.1080/03610918.2019.1577971.
[5] G.A. Bello, M. Arora, C. Austin, M.K. Horton, R.O. Wright, C. Gennings, Extending the distributed Lag Model framework to handle chemical mixtures, Environ. Res. 156 (2017) 253–264, doi:http://dx.doi.org/10.1016/j.envres.2017.03.031.
[6] N. Meinshausen, P. Bühlmann, Stability selection, J. R. Stat. Soc. Ser. B (Statis. Methodol.) 72 (2010) 417–473, doi:http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x.
[7] G. Shmueli, To explain or to predict? Stat. Sci. 25 (2010) 289–310, doi:http://dx.doi.org/10.1214/10-STS330.
[8] T. Yarkoni, J. Westfall, Choosing prediction over explanation in psychology: lessons from machine learning, Perspect. Psychol. Sci. 12 (2017) 1100–1122, doi:http://dx.doi.org/10.1177/1745691617693393.
[9] T. Borovicka, M. Jirina, P. Kordik, M. Jiri, Selecting representative data sets, in: A. Karahoca (Ed.), Advances in Data Mining Knowledge Discovery and Applications, InTech, 2012, doi:http://dx.doi.org/10.5772/50787.
[10] M. Krzywinski, N. Altman, Points of significance: importance of being uncertain, Nat. Methods 10 (2013) 809–810, doi:http://dx.doi.org/10.1038/nmeth.2613.
[11] R Core Team, R: A Language and Environment for Statistical Computing, (2018) .
[12] S. Renzetti, P. Curtin, A.C. Just, G. Bello, C. Gennings, gWQS: Generalized Weighted Quantile Sum Regression [WWW Document]. R Packag. Version 1.1.0 URL, (2018) . https://cran.r-project.org/package=gWQS.
[13] E.M. Tanner, C. Gennings, evamtanner/Repeated_Holdout_WQS: 1st Rodeo (Version v1.0.0). [WWW Document]. Zenodo., (2019), doi:http://dx.doi.org/10.5281/zenodo.2658697.
[14] C.-G. Bornehag, S. Moniruzzaman, M. Larsson, C.B. Lindström, M. Hasselgren, A. Bodin, L.B. von Kobyletzkic, F. Carlstedt, F. Lundin, E. Nånberg, B.A.G. Jönsson, T. Sigsgaard, S. Janson, The SELMA study: a birth cohort study in Sweden following more than 2000 mother-child pairs, Paediatr. Perinat. Epidemiol. 26 (2012) 456–467, doi:http://dx.doi.org/10.1111/j.1365-3016.2012.01314.x.