UNIVERSITÀ DEGLI STUDI DI MILANO
Dottorato di ricerca in "**Epidemiologia, Ambiente e Sanità Pubblica**"
XXXIII Ciclo
Dipartimento di Scienze Cliniche e di Comunità

UNIVERSITÀ DEGLI STUDI DI MILANO

The Weighted Quantile Sum Regression: Extensions and Applications

Settore scientifico disciplinare MED/01

Dottorando: **Stefano RENZETTI** (Matricola: R11883)

Tutor: **Chiar.mo Prof. Monica Ferraroni, Chiar.mo Prof. Stefano Calza**
Coordinatore del Dottorato: **Chiar.mo Prof. Carlo LA VECCHIA**

A.A. 2019/2020

Contents

Abstract

During these last few years an increasing body of scientific evidence showed that looking at the single exposure to chemicals without considering the mixture effect can cause an underestimate of the chemical exposures risk. This poses also statistical challenges on how to manage more complex datasets. Weighted Quantile Sum (WQS) regression is a new statistical model that allows to deal with this problems. It is able to test the association of the overall environmental exposures with an outcome and to find the main actors in the association between the exposure and the dependent variable.
Through this work we showed how we adapted the model to allow to fit a WQS regression in presence of binary, multinomial and count outcomes. Moreover, we implemented two more extensions: the possibility to test for an interaction between the WQS index (representing the overall exposure) and a continuous or categorical variable; and the ability of having two indices in the same model, one looking in the positive and the second in the negative direction when the mixture can have a bidirectional effect on the outcome. The first extension answers to a frequent and important line of inquiry in epidemiologic studies that is whether there is an effect modification (i.e., an interaction) between an exposure and a particular covariate of interest that can affect the association between the exposure and the outcome. The second extension allows to estimate both the protective and harmful effect of the mixture within the same regression model. Lastly, we showed how to apply this novel method in the genetic context thanks to the inclusion of the double WQS index. We then compared its results with the standard methodology used to test the effect of a gene set on a particular phenotype.
The simulation studies performed to test the new extensions showed the good performance of the methods reducing the bias and standard error of the estimates of the effect of the mixture on the outcome and correctly identifying the elements in the mixture that play a major role in the studied association. A high specificity was also observed. Through the case studies we were able to see how WQS confirmed previous major findings and providing new insights respect to previous literature. When we tested for the interaction between age or sex and the exposure to lead (Pb), cadmium (Cd), mercury (Hg), selenium (Se) and manganese (Mn) we found that the association between the forced vital capacity (FVC) and Pb and Hg was attenuated among older children,

while female FVC is more susceptible to Cd and Hg compared to males. The application of the double index to test the association between 43 nutrients and obesity showed a harmful effect of moisture (from all sources), polyunsaturated fatty acids, saturated fatty acids, sodium, caffeine and cholesterol while a protective effect was found for beta-carotene, vitamin B12, vitamin B6, vitamin D, folic acid, vitamin C, folate DFE and alpha-carotene. Finally, through WQS we observed a significant role of the genes involved in cell-cycle in the risk of death for ovarian cancer which was not shown applying single sample Gene Set Enrichment Analysis.

The advantages of WQS regression and the extension that we described in this work are the ease of use and interpretation of the results; moreover, none of the other environmental mixture methods allow to consider the effect modification due to a covariate or to measure the amount of positive and negative association when the elements in the mixture show both effects. This work will be the starting point for additional future extensions, improvements and applications of the model while all these extensions will be implemented in the `gWQS` package of the statistical software *R*.

Chapter 1

Introduction

Humans are daily exposed to multiple chemicals from different sources and the assessment of the effects of the exposure to mixture of chemical contaminants on human health is becoming always more a big concern. During these last few years an increasing body of scientific evidence showed that looking at a single exposure element without considering the mixture effect can cause an underestimate of the chemical exposures risk (Martin et al. 2013, Kortenkamp & Faust 2018, European Commission 2011). The introduction of analytical method yielding multi-elemental measurements in biological samples has further enhanced the potential of biomonitoring in addressing the effects of mixed exposures. With the availability of these information it became of interest to look at the overall impact of the exposure to the mixture and the interactions among chemicals. This poses also statistical challenges on how to manage more complex datasets. Several models with the ability of conducting regression with a set of correlated variables were already available, such as ridge regression (Hoerl & Kennard 1970), lasso (Tibshirani 1996), adaptive lasso (Zou 2006) and elastic net (Zou & Hastie 2005). All these methods are particularly useful for variable selection in prediction models and in particular when the number of covariates is much higher than the number of observations. However, they can have some limitations in assessing the risk of the exposure to chemical mixture: e.g. ridge regression does not reduce the dimensionality while lasso selects a random element among the correlated predictors and elastic net keeps in the model or eliminates all the correlated elements. This is problematic in the context of environmental chemical mixture exposure since we need to find the elements truly associated with the outcome and not because of a "grouping effect". Moreover, none of these methods are able to estimate an overall mixture effect on the outcome

of interest.
New statistical methods were recently developed to address the specific question about measuring the risk of the environmental chemical mixture exposure (Stafoggia et al. 2017). Among others there is the Weighted Quantile Sum (WQS) regression (Czarnota et al. 2015, Gennings et al. 2013, Carrico et al. 2015, Horton et al. 2015, Brunst et al. 2017). This new statistical model constructs a weighted index estimating the mixed effect of all predictor variables on an outcome, which may then be used in a regression model with relevant covariates. WQS is able to test for the association between the overall environmental exposure with a dependent variable or outcome and to find the main actors in the association between the exposure and the dependent variable in a simplistic nevertheless powerful model through the weights estimation. However, WQS in its first formulation can still be improved to better answer to epidemiological questions like testing for the modification effect of a covariate of interest in the association between the exposure mixture and the outcome or to consider the double effect of the mixture on the dependent variable. The first extension will allow to apply WQS also in scenarios where the association between the exposure and the outcome can vary across groups (e.g. sex) or at different levels of a continuous variable (e.g. age) which is often the case in epidemiological studies. The introduction of an interaction term is feasible in WQS regression thanks to the WQS index which we can treat as a continuous variable in a classical regression once estimated by the model. The additional ability to test for both positive and negative effect of the mixture on the dependent variable with the estimate of a double index, will allow to apply WQS in context where the mixture can have a bidirectional association (e.g. nutrients) with the outcome. The inclusion of two indices was already possible in the original WQS formulation but not at the step where the weights of the index are estimated. Through this work we will show how to deal with a double index when the model estimates the weights to be attributed to each element in the mixture. This second extension will also give the possibility to apply WQS in different context like genetic which is characterized by the bidirectional effect of the expression of genes on a phenotype.
In the next chapters we are going to show how the weights and the parameters are estimated in the WQS regression and how we adapted the model to allow for logistic, multinomial, Poisson and negative binomial regression. We

will then show two improvements of the model: the possibility to test for an interaction between the WQS index (representing the overall exposure) and a continuous or categorical variable; and the ability of having two indices in the same model, one looking in the positive and the second in the negative direction when the mixture is made of both protective and harmful elements. Lastly, we applied this last method in the context of gene expression and its effect on a phenotype of interest. All these extensions of the WQS regression were tested on simulated data and then applied in case studies to compare the results with previous literature.

Chapter 2

The Weighted Quantile Sum (WQS) Regression

2.1 Overview of the method

As previously mentioned, WQS regression is a statistical model for multivariate regression in high-dimensional datasets commonly encountered in environmental exposures, epi/genomics, and metabolomic studies, among others. The method is divided in two parts: a training step where the weights are estimated and a validation step where the regression is fitted using the previously estimated weights. The WQS model has the following equation:

$$g(\mu) = \beta_0 + \beta_1 \left(\sum_{i=1}^{c} w_i q_i \right) + \boldsymbol{z'\varphi} \tag{2.1}$$

where g is the link function as in generalized linear model, μ is the mean of the outcome, q_i is the quantile of the i^{th} component, w_i is the weight (to be estimated) associated with the i^{th} component, $\boldsymbol{z'}$ is the vector of covariates and $\boldsymbol{\varphi}$ is the vector of parameters associated with the covariates. The $(\sum_{i=1}^{c} w_i q_i)$ term represents the index that weights and sums the components included in the mixture. To estimate the model, the data may be split in a training and a validation dataset: the first one to be used for the weight estimation, the second one to test for the significance of the final WQS index. The weights are estimated through a bootstrap and constrained to sum to one and bounded between zero and one: $\sum_{i=1}^{c} w_i = 1$ and $0 \leq w_i \leq 1$. For each bootstrap sample (usually $B = 100$ total samples) a dataset is created sampling with replacement from the training dataset and the parameters of the model in equation 2.1 ($\theta = (\beta_0, \beta_1, w_1, \ldots, w_c, \boldsymbol{\varphi})$) are estimated through

an optimization algorithm where the log-likelihood is the objective function:

$$\hat{\theta}_{WQS} = \underset{\theta}{\operatorname{argmax}} \left[l(\theta; y) + \lambda \left(\sum_{i=1}^{c} w_i - 1 \right) \right] \tag{2.2}$$

where $l(\theta; y)$ is the log-likelihood function and λ is the lagrangian coefficient associated with the equality constraint in which the weights have to sum to 1. An inequality constraint is also applied in order to impose that $0 \leq w_i \leq 1$.
Once the weights are estimated the model is fitted in order to find the regression coefficients in each ensemble step. After the bootstrap ensemble is completed, the estimated weights are averaged across bootstrap samples to obtain the WQS index:

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

where $\bar{w}_i = \frac{1}{\sum_{b=1}^{B} f(\beta_{1(b)})} \sum_{b=1}^{B} w_{i(b)} f(\beta_{1(b)})$ and $f(\beta_{1(b)})$ is a signal function that we will specify later in the text. Typically weights are estimated in a training set then used to construct a WQS index in a validation set, which can be used to test for the association between the mixture and the health outcome in a standard generalized linear model, as:

$$g(\mu) = \beta_0 + \beta_1 WQS + \boldsymbol{z'\varphi}$$

Due to the structure of the model either a positive or a negative direction of the association between the dependent variable and the WQS index has to be chosen; that is, the model is inherently one-directional, in that it tests only for mixture effects positively or negatively associated with a given outcome. In practice analyses should therefore be run twice to test for associations in either direction. The specification of a test for positive or negative association determines the form of the signal function:

$$f(\hat{\beta}_{1(b)}) = \begin{cases} 1, & \text{if } \hat{\beta}_{1(b)} \text{ and the chosen direction have the same sign} \\ 0, & \text{if } \hat{\beta}_{1(b)} \text{ and the chosen direction have different sign} \end{cases}$$

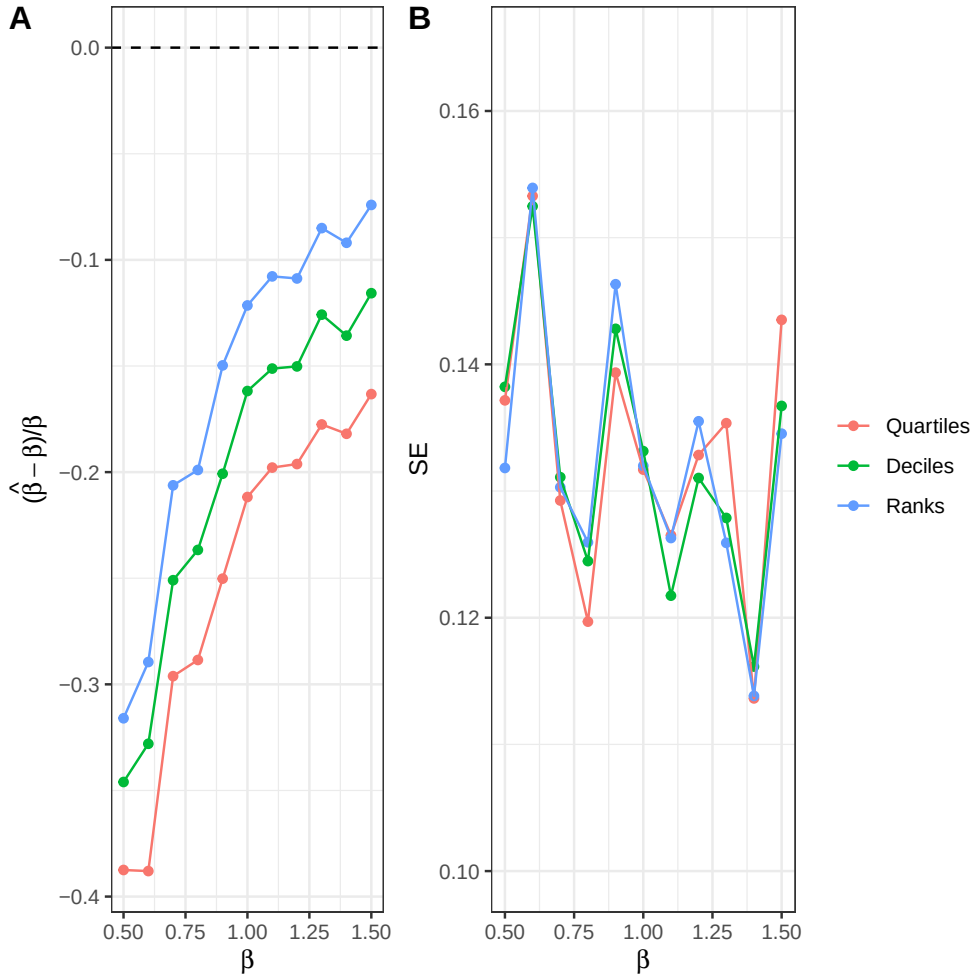
The one-directional index allows not to incur in the reversal paradox when we have highly correlated variables (Tu et al. 2008), moreover the bootstrap

step (or random subsets of components as described below in section 2.4) improves the identification of bad actors.
After the final model is fitted we can test the significance of the β_1 to see if there is an association between the WQS index and the outcome. In the case the coefficient is significantly different from 0 then we can interpret the weights: the highest values identify the associated components as the relevant contributors in the association. A selection threshold can be decided a priori as $\tau = 1/c$ to identify those chemicals that have a significant weight in the index.
Once the WQS index is estimated, the required assumptions for generalized linear regression has to be met.

2.2 Quantiles choice

Quantiles are applied to get measure of the variables included in the mixture on the same scale. We simulated data to better understand how the choice of quartiles, deciles and ranks affect the estimate of the regression parameter associated to the WQS index in terms of bias and standard error. We considered 25 exposure concentrations simulated from a distribution of phthalate biomarkers measured in subjects participating in the NHANES study (2001-2002) (NCHS 2017). Four elements were randomly selected and were assigned a non null weight as follows: $w_{15} = 0.5$, $w_{19} = 0.25$, $w_{14} = 0.15$ and $w_3 = 0.1$. We then generated 11 different dependent variables from a normal distribution with standard deviation equal to two and mean equal to the combination of parameters and variables as per equation 2.1 setting 11 different values of the regression parameter β: $\beta = 0.5$, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4 and 1.5. Each WQS index was standardized to get comparable effect estimates when using the different quantiles before the dependent variables were generated. A WQS regression using quartiles, deciles or ranks was applied to estimate the effect of the mixture on each outcome. Figure 2.1A shows that all the three models reduced the bias at increasing values of β and ranks usage produced the lowest bias compared to quantiles and deciles in all different scenarios. The Standard Errors (SE) were similar among the three different models fluctuating around 0.13 and not showing a trend at increasing β (figure 2.1B).

Figure 2.1: Relative bias and standard error (SE) of the estimate of the regression parameter associated to the WQS index (β) at varying values of β when using quartiles, deciles or ranks to stnadardize the mixture components.



Based on these results we suggest to apply ranks as a way to scale the elements included in the mixture. Standardization can also be applied, however the estimates will be more sensitive to extreme values.

2.3 The Generalised WQS regression

The WQS regression can be generalised and applied to multiple types of dependent variables. To adapt the model to different types of outcomes we need to specify the objective function reported in equation 2.2 defining different log-likelihoods depending on the outcome distribution. This allows to estimate the weights taking into account the different distribution of the

dependent variable; once we are able to estimate the WQS index we can fit a standard generalised linear model. In particular we adapted the WQS regression to four different cases: logistic, multinomial, Poisson and negative binomial regression. For these last two cases we also added the possibility to fit zero-inflated models keeping the same objective function used to estimate the weights as for the Poisson and negative binomial regression but taking into account the zero inflation fitting the final model.
Starting from a linear regression the following function is minimised when estimating the weights:

$$\begin{aligned}\hat{\theta}_{WQS} &= \underset{\theta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \boldsymbol{z'\varphi} \right) \right)^2 \right. \\ &\quad \left. + \lambda \left(\sum_{j=1}^{c} w_j - 1 \right) \right]\end{aligned}$$

For a logistic regression the following likelihood is maximised:

$$\begin{aligned}\hat{\theta}_{WQS} &= \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \left(y_i \times \log \left(\frac{1}{1 + \exp(\beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \boldsymbol{z'\varphi})} \right) \right. \right. \\ &\quad \left. \left. + \left(1 - y_i \right) \times \log \left(1 - \beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \boldsymbol{z'\varphi} \right) \right] \right.\end{aligned}$$

The equation to be maximised for a multinomial regression is the following:

$$\begin{aligned}\hat{\theta}_{WQS} &= \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \left[\sum_{l=1}^{L-1} \left(y_{ij} \left(\beta_{0l} + \beta_{1l} \sum_{j=1}^{c} w_{lj} q_{ij} + \boldsymbol{z'\varphi} \right) \right. \right. \right. \\ &\quad \left. \left. \left. - \log \left(1 + \sum_{l=1}^{L-1} \exp \left(\beta_{0l} + \beta_{1l} \sum_{j=1}^{c} w_{lj} q_{ij} + \boldsymbol{z'\varphi} \right) \right) \right) \right] \right\}\end{aligned}$$

The objective function used to estimate the weights in a Poisson regression is:

$$\hat{\theta}_{WQS} = \underset{\theta}{\text{argmax}} \left[\sum_{i=1}^{n} \left(y_i \times \left(\beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \boldsymbol{z'\varphi} \right) - \exp\left(\beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \boldsymbol{z'\varphi} \right) \right) \right]$$

In the case of a negative binomial regression the likelihoods to be maximised is:

$$\hat{\theta}_{WQS} = \underset{\theta}{\text{argmax}} \left[\sum_{i=1}^{n} \left(y_i \log(\alpha) + y_i \left(\beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \boldsymbol{z'\varphi} \right) - \left(y_i + 1/\alpha \right) \log\left(1 + \alpha \exp\left(\beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \boldsymbol{z'\varphi} \right) \right) + \log(\Gamma(y_i + 1/\alpha)) - \log(\Gamma(y_i + 1)) - \log(\Gamma(1/\alpha)) \right) \right]$$

All these adaptations of the WQS regression to different type of outcomes were implemented in the *R* package `gWQS` as shown in Appendix A.

2.4 A Random Subset Implementation of Weighted Quantile Sum ($\mathbf{WQS_{RS}}$) Regression

A novel implementation of WQS regression for high-dimensional mixtures with highly correlated components was proposed in Curtin et al. (2019). This approach to which we will refer as WQS_{RS} applies a random selection of a subset of the variables included in the mixture instead of the bootstrapping for parameter estimation. Through this method we are able to generate a more de-correlated subsets of variables and reduce the variance of the parameter estimates compared to a single analysis.
In this case the formula that describes how the weights are averaged after their estimation for each subset ($\boldsymbol{w}_{b,(s\times 1)}$) to obtain the full ensemble set $\overline{\boldsymbol{w}}_{c\times 1}^{FULL}$ is the following:

$$\overline{\boldsymbol{w}}^{FULL}_{c\times 1} = \frac{1}{\sum_{b=1}^{B} f(\beta_{1(b)})} \sum_{b=1}^{B} (\boldsymbol{A}_b)'_{(c\times s)} \boldsymbol{w}_{b,(s\times 1)}$$

where s is the number of randomly selected variables (say, $s = \sqrt{c}$), $\boldsymbol{A}_b$ is the translation matrix with elements $a_{q,j} = \{f(\beta_{1(b)}), j \in H\}; q = 1, ..., s; H \in \{1, 2, ..., c\}$ and $f(\beta_{1(b)})$ is the same signal function already seen for the bootstrap WQS that we will specify later in the text.
This novel statistical methodology was shown to be more effective compared to WQS in modeling contexts with large predictor sets, complex correlation structures, or where the numbers of predictors exceeds the number of subjects.

2.5 Repeated holdout validation for weighted quantile sum regression

One limit of WQS is the reduced statistical power caused by the necessity to split the dataset in training and validation sets. This partition can also lead to unrepresentative sets of data and unstable parameter estimates. A recent work from Tanner et al. (2019) showed that conducing a WQS on the full dataset without splitting in training and validation produces optimistic results and proposed to apply a repeated holdout validation combining cross-validation and bootstrap resampling. From now on we will refer to this approach as WQS_{RH}. They suggested to repeatedly split the data 100 times with replacement and fit a WQS regression on each partitioned dataset. Through this procedure we obtain an approximately normal distribution of the weights and the regression parameters and we can apply the mean or the median to estimate the final parameters. We can then use their distributions to build the 95% confidence intervals (CI) based on the standard deviation or using the 2.5^{th} and 97.5^{th} percentiles. Another advantage of the WQS_{RH} is the ability to characterize the weight distribution. On the other hand a limit of this approach is the higher computational intensity, for this reason only 100 repeated holdout validation iterations were proposed in Tanner et al. (2019), but if feasible, a larger number of repetitions would allow to better meet a normal approximation (usually ≥ 1000 for bootstrap).

The *R* package `gWQS` was developed to make this new method available and it is downloadable from CRAN repository. The Appendix A shows how to use the package through four different applications of the model on a simulated dataset.

Chapter 3

Interaction between WQS index and a continuous or a categorical variable

A frequent and important line of inquiry in epidemiologic studies is whether there is effect modification (i.e., an interaction) between an exposure and a particular covariate of interest that can affect the association between the exposure and the outcome (Knottnerus & Tugwell 2019, Stafoggia et al. 2017). Traditional modelling strategies, particularly forms of generalized linear regression, can be easily applied to capture these effects through the estimation of multiplicative effects; for example, one might test for sex-based modification of exposure effects by estimating the interaction between sex- and exposure-based effects. The increasing availability of high-dimensional exposure assessments, however, has compounded the challenge in estimating the effects associated with individual exposure variables, as these must ideally be evaluated in the context of combined exposures.

Here we propose a method to extend the generalized mixtures modeling strategy advanced with WQS regression to the evaluation of effect modification by covariates. This method retains the advantages of a mixtures modeling strategy, in that it allows evaluation of the association between multiple combined exposures and an outcome rather than evaluating effects in discrete models, but also considers how the interaction between the whole mixture exposure and the covariate affects the outcome.

To demonstrate the utility and efficacy of this procedure, we show how we extended WQS regression to test for an interaction between the WQS index, representing the mixture effect, and a covariate of interest, with weights estimated in the presence of the interaction. Even within a consistent modeling framework, there are multiple approaches to estimating multiplicative effects.

For example, interactions could be considered during or after ensemble estimation and aggregation, or at different stages of cross-validation. As such we compared the bias introduced in implementing several different approaches to estimating effect moderation. In particular we will show that when the interaction term is kept in the model both during the weight estimation step and when fitting the final model on the validation dataset, WQS shows much better performance in both estimating the mixture effect and identifying the "main actors". We characterize the efficacy of this procedure in contexts of interaction with a continuous variable and with a categorical variable. We then applied this new extension of WQS to a real case study from Madrigal et al. 2018 (Madrigal et al. 2018) considering data from the National Health and Nutrition Examination Survey (NHANES) (2011-2012) where the association between heavy metal exposure and pulmonary function among children and adolescents was tested.

3.1 Model and Methods

The standard formulation of WQS regression (equation 2.1) can be extended to evaluate interactions between the WQS index and a continuous covariate, x. The WQS general formula is of the form:

$$g(\mu) = \beta_0 + \beta_1 \left(\sum_{i=1}^{c} w_i q_i \right) + \beta_2 x + \beta_3 x \sum_{i=1}^{c} w_i q_i + \boldsymbol{z'}\boldsymbol{\varphi} \tag{3.1}$$

where g is the link function as in generalized linear model, μ is the mean of the outcome, q_i is the quantile of the i^{th} component, w_i is the weight (to be estimated) associated with the i^{th} component, $\boldsymbol{z'}$ is the vector of covariates and $\boldsymbol{\varphi}$ is the vector of parameters associated with the covariates. As we can see comparing equation 2.1 with equation 3.1, the interaction term is incorporated and identified by the term $x \sum_{i=1}^{c} w_i q_i$ between the WQS index and the continuous variable x. The interaction parameter is adjusted for both during the training step where the weights are estimated and in the final model in the holdout validation data using the constructed index. Here, the interaction model allows for a change of slope associated with the WQS index at fixed levels of x, say $x0$, the change of slope is $\beta_3 x0$, and for a change of intercept is given by $\beta_2 x0$.

WQS can similarly evaluate interactions involving categorical covariates. In this case, the β_1 parameter and weights vary across the levels of the categorical variable as shown by the formula:

$$\begin{aligned} g(\mu) = \beta_0 + \beta_1 \left(\sum_{i=1}^{p} \sum_{j=1}^{c} w_{ij} q_{cat_{ij}} \right) + \beta_{21} x_{11} + ... + \beta_{2(p-1)} x_{1(p-1)} + \\ \beta_{31} x_{11} \left(\sum_{j=1}^{p-1} \sum_{i=1}^{c} w_{ij} q_{cat_{ij}} \right) + ... + \\ \beta_{3(p-1)} x_{1(p-1)} \left(\sum_{j=1}^{p-1} \sum_{i=1}^{c} w_{ij} q_{cat_{ij}} \right) + \boldsymbol{z'\varphi} \end{aligned} \tag{3.2}$$

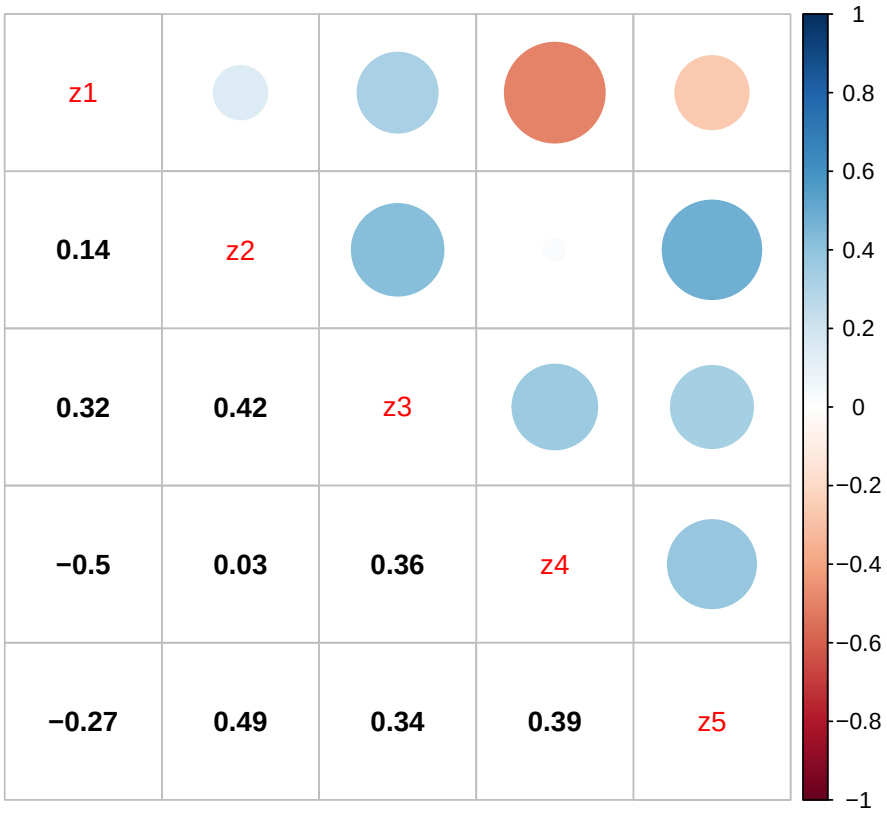
where $x_{11}, ..., x_{1(p-1)}$ are the dummy variables of the categorical variable x_1 with p categories (x_{10} considered as reference category) and $q_{cat_{ij}}$ are the quantile variables associated to each mixture element ($i = 1, ..., c$) that are equal to zero if the considered observation belongs to a level of x_1 which is different from the one examined. For example, if x_1 has two categories A and B then $q_{cat_{iA}} = \begin{cases} q_i, & \text{if } x_1 = A \\ 0, & \text{otherwise} \end{cases}$ and $q_{cat_{iB}} = \begin{cases} q_i, & \text{if } x_1 = B \\ 0, & \text{otherwise} \end{cases}$ as shown in Brunst et al. (2017) (Brunst et al. 2017). In this case w_{ij} represents the weight associated with the i^{th} component for the j^{th} category and the following constraints are applied $\sum_{i=1}^{c} w_{ij} = 1$ for $j = 1, .., p$; $0 \leq w_{ij} \leq 1$ for $i = 1, ..., c$; $j = 1, ..., p$. In this case, the interaction model allows for different slopes for the WQS index and for different intercepts for each category while also allowing different weights for each category.

3.1.1 Simulation Study

In order to test the goodness of fit of the WQS regression in the presence of an interaction between the WQS index and a continuous or categorical variable we performed a simulation study. The data were taken from the blood metal concentrations of the 2011-2012 NHANES cohort. In particular lead (Pb), cadmium (Cd), mercury (Hg), selenium (Se) and manganese (Mn) were considered. In total 100 different datasets were generated from a multivariate normal distribution with 1000 observations and null mean. The

variance-covariance matrix of the five variables was applied to reproduce their correlation structure. Figure 3.1 shows a complex correlation matrix where we can see a wide range of correlations among all the five variables (from -0.5 to 0.49).

Figure 3.1: Spearman correlation matrix of the five heavy metals (lead, cadmium, mercury, selenium and manganese) taken from the 2011-2012 NHANES study.



Five dependent variables were then generated from a normal distribution with unit variance and mean equal to the resulting WQS formula setting the parameters as described in table 3.1. All the remaining weights associated to the remaining elements in the mixture not displayed in table 3.1 were set to zero.

The outcome y_0 was generated in the absence of an interaction with either a continuous or a categorical covariate to check the specificity of the models. The dependent variable y_1 was built in the presence of an interaction be-

Table 3.1: Regression parameter (β_i) and weight (W_i) values used to generate the dependent variables in five different models.

	y_0: No interaction	y_1: Interaction with continuous modifier	y_{21}: Interaction with categorical modifier, equivalent weights across categories	y_{22}: Different weights across categories, with no interaction	y_{23}: Interaction with categorical modifier, different weights across categories
β_1	0.5	0.2	0.3	0.5	0.3
β_2	0.3	0.4	0	0	0
β_3	0	0.6	0.3		0.3
W_2	0.3	0.3	0.3		
W_3	0.5	0.5	0.5		
W_5	0.2	0.2	0.2		
W_{2A}				0.6	0.6
W_{5A}				0.4	0.4
W_{1B}				0.3	0.3
W_{3B}				0.7	0.7

tween the WQS index and a continuous variable, while y_{21}, y_{22} and y_{23} were defined in the presence of an interaction with a categorical variable: y_{21} did not account for varying weights across variable categories, y_{22} did not consider the interaction term between WQS index and the categorical covariate but included different weights depending by the categorical variable levels while both interaction term and varying weights were considered to generate y_{23}. Finally, a continuous variable was generated from a standard normal distribution while a categorical variable was generated from a Bernoulli with probability equal to 0.5 and the two categories A and B were defined.
To test the performance of the novel method in the case of an interaction between the WQS index and a continuous covariate (method 2) we compared it to a model where we consider the interaction term only in the validation step (method 1). In the case of an interaction with a categorical variable the first method included only the interaction term between the WQS index and the covariate (method 1); method 2 only considered stratified weights by category while method 3 comprised both the interaction term and the stratified weights.

3.1.2 Case Study

We then applied this method to an empirical case study. A detailed description of the study population and the outcome and exposure variables were reported in NHNAES documentation (CDC 2011, NCHS 2017) and Madri-

gal et al. (2018). Briefly, a total of 1,234 subjects between 6 to 17 years old from the NHANES 2011-2012 survey cycle were included in the study. Exclusion criteria applied for when the spirometry test included chest pain at the time of the exam; recent surgery of the eye, chest or abdomen; tuberculosis exposure; a physical problem with forceful expiration; and anyone with a recent incident of cough with blood or painful ear infections. The recommendations of the American Thoracic Society (ATS) (Miller et al. 2005) were followed for the spirometry measurements. In this application we only considered the Force Vital Capacity (FVC) and we included in the analysis values rated A (exceeds ATS data collection standards) or B (meets ATS data collection standards). Pb, Cd, Hg, Se and Mn were measured in blood using inductively coupled plasma mass spectrometry and ranked in deciles in the WQS regression. The interaction between the WQS index and age and sex was tested applying the method described above. Following Madrigal et al. (2018), the regression models were adjusted for race (non-Hispanic white, non-Hispanic black, Mexican American, other Hispanic, or other/multiracial), the ratio of family income to poverty (using the Department of Health and Human Services (HHS) poverty guidelines), serum cotinine levels, BMI and use of anti-asthmatic, bronchodilator, or inhaler. FVC, serum cotinine and BMI were log-transformed when included in the regression because of their asymmetric distribution.

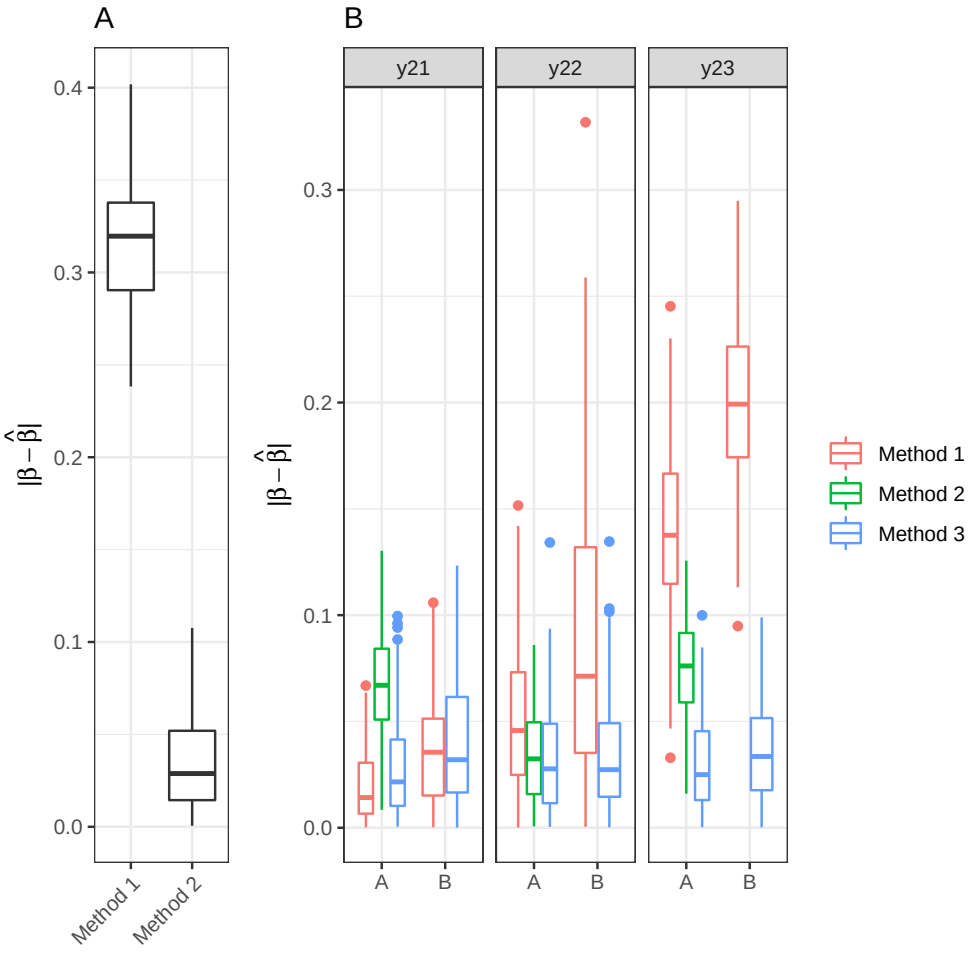
3.2 Results

3.2.1 Simulation Studies

We first examined the case of the interaction between the WQS index and a continuous variable. Figure 3.2A shows the box-plots of the absolute difference between the 100 estimates of the parameter β_3 related to the interaction term and the true value. As we can see from figure 3.2A method 2 gives more accurate measures of the β_3 compared to method 1 (mean error (ME): 0.035; SD: 0.026 and ME: 0.317; SD: 0.040 respectively) reducing the ME estimates of 89.0%. When we consider the interaction with a categorical variable we fit the three different models described in the methods section for each of the three outcomes y_{21}, y_{22} and y_{23}. Figure 3.2B shows that method 1 provides good estimates for the outcome y_{21} but shows a higher bias when considering y_{22} and y_{23} for both level A (ME: 0.029; SD: 0.026, ME: 0.052; SD: 0.036

and ME: 0.129; SD: 0.039 respectively) and level B (ME: 0.037; SD: 0.026, ME: 0.093; SD: 0.073 and ME: 0.194; SD: 0.042 respectively). Method 2 shows better performances when y_{22} is the outcome (ME: 0.033; SD: 0.022) but with a higher bias when y_{21} (ME: 0.068; SD: 0.025) and y_{23} (ME: 0.074; SD: 0.024) are the dependent variables. Method 3 provides good estimates for the interaction coefficient in all three cases for both level A (ME: 0.029; SD: 0.026, ME: 0.029; SD: 0.023 and ME: 0.034; SD: 0.025 for y_{21}, y_{22} and y_{23} respectively) and level B (ME: 0.042; SD: 0.031, ME: 0.042; SD: 0.034 and ME: 0.040; SD: 0.027 for y_{21}, y_{22} and y_{23} respectively). In particular a better performance is shown for y_{23} where we can see a reduction of ME of the 73.6% and 79.4% for level A and B respectively compared to method 1 and the 54.1% compared to method 2. For all the following figures including boxplots we will interpret a more accurate parameter estimation those boxplots showing distributions closer to zero.

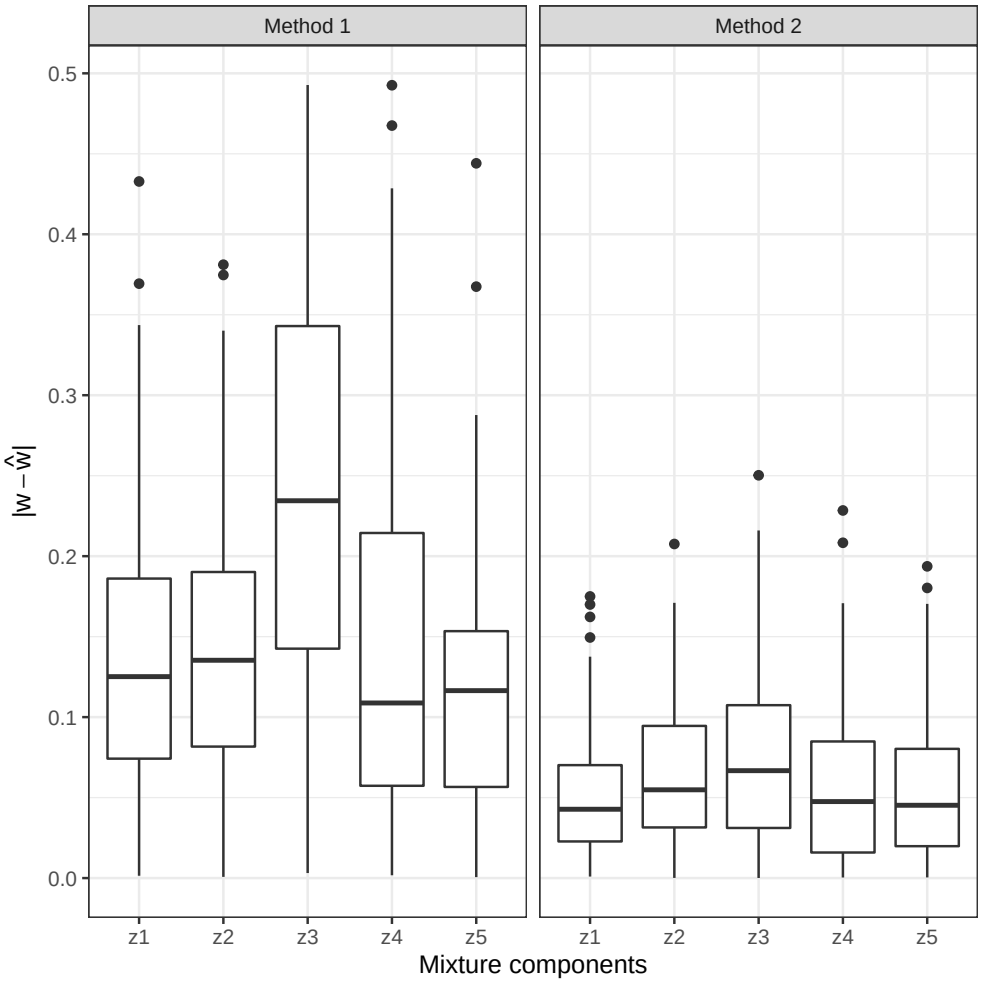
Figure 3.2: Absolute difference between the 100 estimates of the parameter β_3 associated with the interaction term between the WQS index and the continuous covariate x (A) or the categorical variable x_{cat} (B) obtained applying the different methods: in panel (A) method 1 included the interaction only in the validation step while method 2 also included the interaction term when estimating the weights; in panel (B) method 1 included only the interaction term without stratified weights, method 2 only considered stratified weights by category without an interaction term, while method 3 comprised both the interaction term and the stratified weights. The three methods were all applied in the three different scenarios y_{21} (interaction with categorical modifier, equivalent weights across categories), y_{22} (different weights across categories, with no interaction) and y_{23} (interaction with categorical modifier, and different weights across categories). In this second panel the parameters for both levels A and B of x_{cat} are shown.



In figure 3.3 the absolute differences between the estimates of each mixture component weight and the true values in the presence of the interaction between the WQS index and the continuous variable are represented. Method 2 clearly shows estimates of the weights (ME: 0.052; SD: 0.040, ME: 0.063; SD: 0.042, ME: 0.079; SD: 0.058, ME: 0.056; SD: 0.049, ME: 0.055; SD: 0.040 for z1, z2, z3, z4 and z5 respectively) have less absolute bias for all the elements in the mixture compared to method 1 (ME: 0.132; SD: 0.092, ME: 0.136; SD:

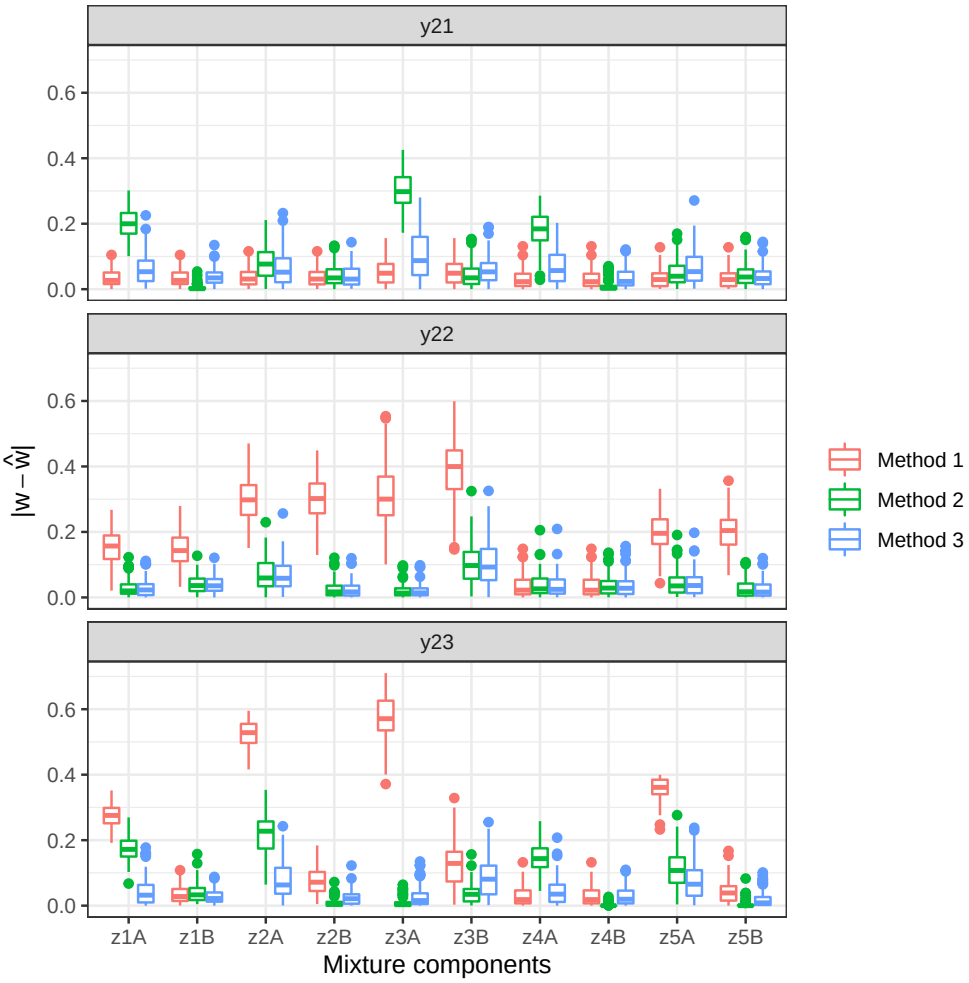
0.085, ME: 0.238; SD: 0.132, ME: 0.139; SD: 0.111, ME: 0.114; SD: 0.075 for z1, z2, z3, z4 and z5 respectively).

Figure 3.3: Boxplots of the absolute difference between the mixture component weight estimated on each one of the 100 datasets by both method 1 and 2 and the true value of each element in the mixture. Method 1 included the interaction only in the validation step while method 2 also included the interaction term when estimating the weights.



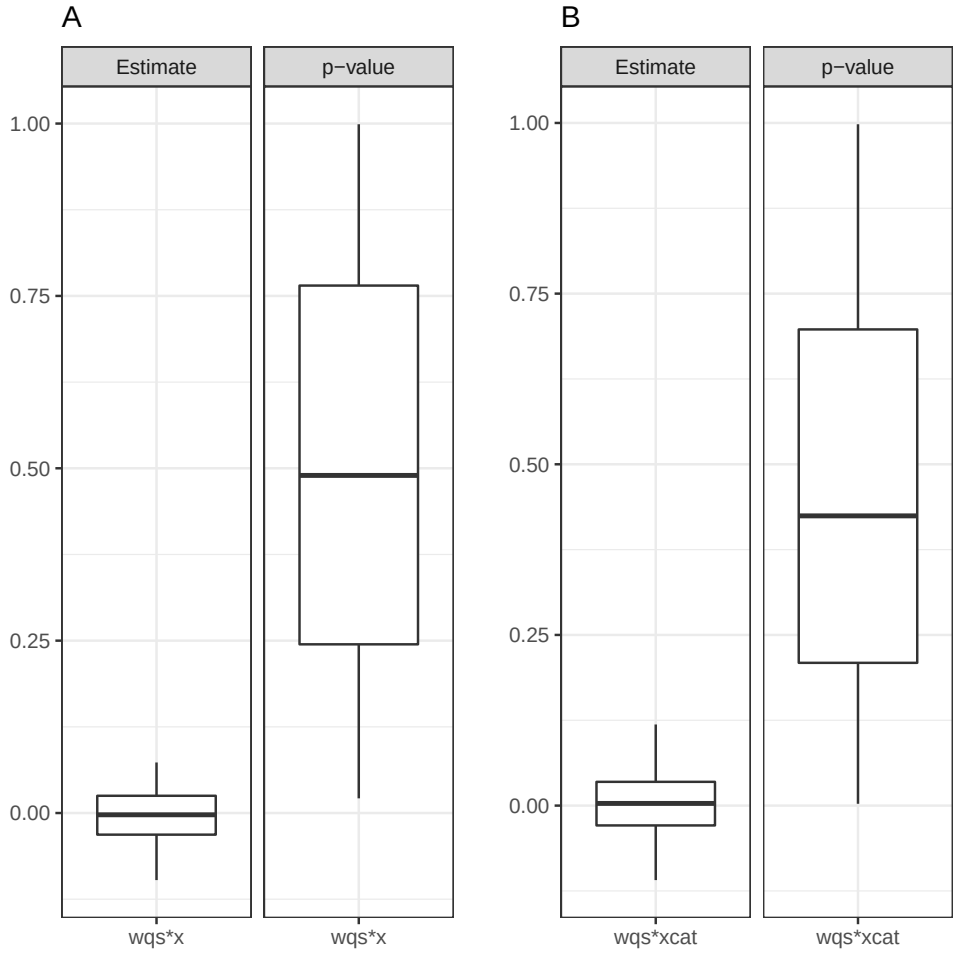
The same analysis was performed on the interaction of the WQS index with the categorical variable. Figure 3.4 shows the absolute difference between the estimated weights and the true value for each element by categorical strata. As for the regression parameter estimates we can still see a more accurate weight estimation of method 3 in all the three different scenarios while method one and method two showed lower bias in the weight estimates when considering y_{21} and y_{23} respectively.

Figure 3.4: Boxplots of the absolute difference between the 100 estimates and the true weight value for each element of the mixture and each strata of the categorical variable. Results are shown for the three methods applied in the three different scenarios depending on the outcome variables y_{21}, y_{22} and y_{23}.



We then checked the specificity of the method fitting a WQS regression including the interaction term with a continuous or a categorical variable when the dependent variable was generated in the absence of any interaction. Figure 3.5 represents the regression parameter estimates of the interaction term between the WQS index and both the continuous (figure 3.5A) and categorical (figure 3.5B) variables and their p-values. We can see that the estimates of parameter β are close to zero in both cases and they have high p-values with only 4% and 5% false positives (p-value below 0.05) for the continuous and categorical variable respectively.

Figure 3.5: Boxplots of the parameter estimates and associated p-values of the interaction term between the WQS index and the continuous (A) and categorical (B) variables.



3.2.2 Case Study

Table 3.2 shows the descriptive statistics of the variables included in the model. The final cohort was composed of 1,234 subjects with an average age of 11.5 ranging from 6 to 17 years old, a slightly higher percentages of males (50.4%) and a higher proportion of non-Hispanic black and non-Hispanic white.

We considered three models: WQS regression without interaction; WQS regression allowing for a $WQS \times age$ interaction; and stratified WQS regression with interaction with sex. In all three models, we find that there is a significant reduction of FVC at higher levels of exposure to metals (higher values of the WQS index) (table 3.3). In particular the effect is mainly driven by Hg, Pb and Mn which are the mixture components with a higher weight

Table 3.2: Descriptive statistics of the variables included in the WQS regression.

	Overall (N=1234)
FVC (ml)	
Mean (SD)	2880.4 (1062.4)
Age	
Mean (SD)	11.5 (3.3)
Sex	
M	622 (50.4%)
F	612 (49.6%)
Family income to poverty ratio	
Mean (SD)	2.1 (1.6)
Race	
Mexican American	237 (19.2%)
Other Hispanic	132 (10.7%)
Non-Hispanic White	305 (24.7%)
Non-Hispanic Black	355 (28.8%)
Other Race	205 (16.6%)
Use of antiasthmatic, bronchodilator, or inhaler	
No	1114 (90.3%)
Yes	120 (9.7%)
BMI (kg/m2)	
Mean (SD)	21.2 (5.6)
Serum cotinine (ng/mL)	
Mean (SD)	3.8 (26.9)
Blood Cd (ug/L)	
Mean (SD)	0.2 (0.1)
Blood Mn (ug/L)	
Mean (SD)	10.6 (3.6)
Blood Pb (ug/L)	
Mean (SD)	0.8 (0.6)
Blood Se (ug/L)	
Mean (SD)	181.3 (21.8)
Blood Hg (ug/L)	
Mean (SD)	0.6 (0.7)

(figure 3.7A); this means that if we move from a decile to the next one of Hg, Pb and Mn distribution we can observe a decrease of $-0.02 \times 0.393 = -0.008$, $-0.02 \times 0.243 = -0.005$ and $-0.02 \times 0.199 = -0.004$ ($\beta \times w$) in log-transformed FVC on average respectively. When we included the interaction between the WQS index and child age in the model we see that there is a positive significant interaction meaning that the negative effect on FVC of the mixture diminishes at increasing age (table 3.3 and figure 3.6A). The elements that mainly show this pattern are Pb and Hg (figure 3.7B). The interaction between the WQS index and sex was finally considered in the regression. A marginally significant negative effect of the mixture was found among males (with slope estimate -0.01; p=0.052) while females were significantly (p=0.010) more susceptible to metals (with slope estimate -0.04=-0.01-0.03; table 3.3 and figure 3.6B). When looking at the weights associated with each metal we can further see that the single elements have different effects on the outcome depending on sex: Cd and Hg were more toxic among females while Mn and Hg had a higher impact on males FVC (figure 3.7C).

Table 3.3: Regression coefficients, 95% confidence intervals (CI) and p-values from the three WQS regression where no interaction, the interaction with age and with sex were considered to test the association between metal exposure and FVC. All regression models were adjusted for race, the ratio of family income to poverty, serum cotinine levels and BMI.

	No interaction		Interaction with age		Interaction with sex	
	β (95%CI)	p-value	β (95%CI)	p-value	β (95%CI)	p-value
WQS	-0.02 (-0.03, -0.01)	<0.001	-0.02 (-0.04,-0.01)	0.002	-0.01 (-0.02, 0.0001)	0.052
Age	0.27 (0.26, 0.29)	<0.001	0.27 (0.26, 0.28)	<0.001	0.28 (0.26, 0.29)	<0.001
Sex F vs M	-0.12 (-0.15,-0.10)	<0.001	-0.13 (-0.15,-0.10)	<0.001	-0.08 (-0.14,-0.02)	0.007
WQS*Age			0.02 (0.01, 0.03)	0.008		
WQS*Sex F vs M					-0.03 (-0.05,-0.01)	0.010

Figure 3.6: Trends representation between the WQS index and the FVC (log-transformed) at varying age levels (10^{th}, 25^{th}, 40^{th}, 50^{th}, 60^{th}, 75^{th} and 90^{th} percentiles were used) when the interaction with age was included in the model (panel A) and for males and females when the interaction with sex was considered (panel B).

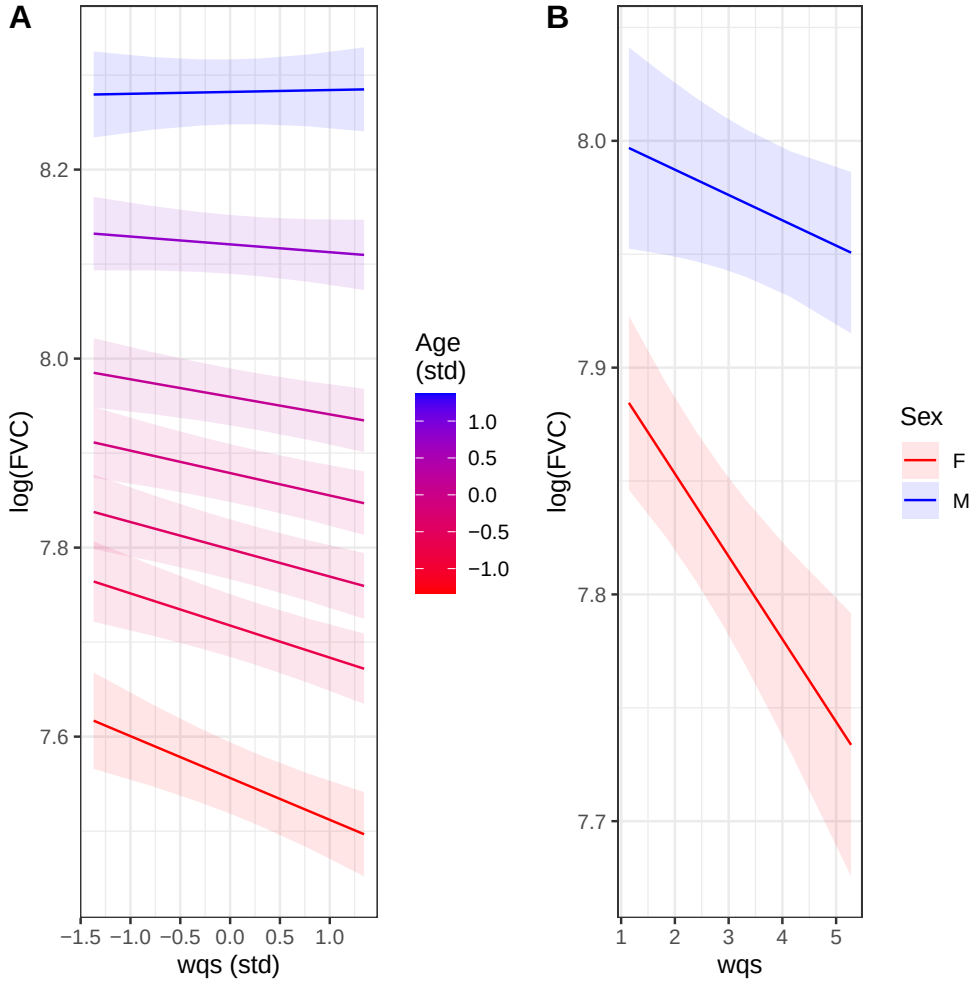
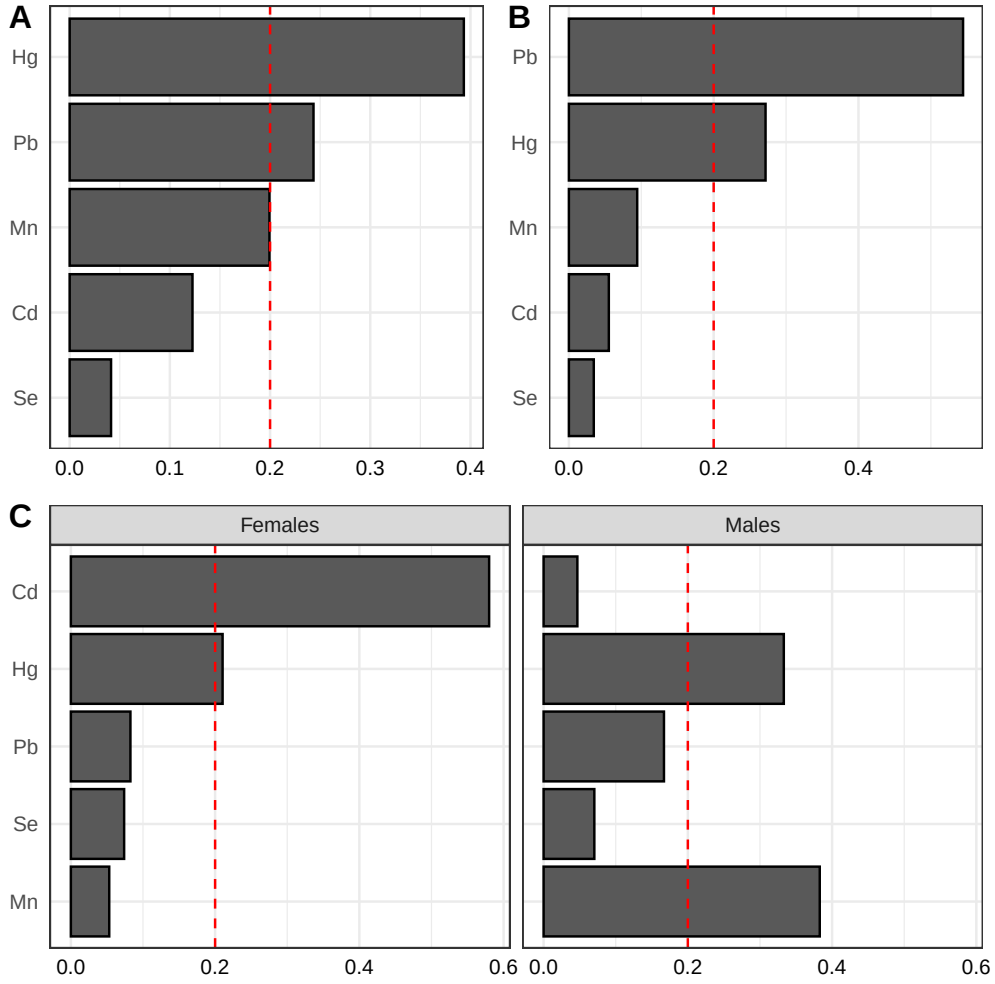


Figure 3.7: Bar plot of the weights associated with the metals included in the mixture estimated in the three WQS regression models where no interaction (A), the interaction with age (B) and sex (C) were considered.



Chapter 4

A Weighted Quantile Sum Regression with Double Index

As described in chapter 2, Weighted Quantile Sum (WQS) regression is a recent statistical model that is increasingly applied in epidemiological studies to solve problems like multiple comparisons and multicollinearity that are typical of situations where we have high dimensional and correlated exposures. This method allows to build an empirical weighted index that represents the overall mixture exposure and test the association with the outcome of interest. The body burden index reduces the dimensionality and it is more robust to multicollinearity (Carrico et al. 2015). The original methodology provides the estimate of a single index that allows to measure the association between the mixture and the dependent variable in only one direction (either positive or negative). This can be an advantage when the elements in the mixture have the same direction in the association with the dependent variable. In fact, looking in one direction we avoid to incur in the reversal paradox (Tu et al. 2008). However, when the mixture is made of both "good" and "bad" actors, this can become a limit when we want to estimate both the positive and negative effect that the mixture has on the specific outcome of interest. In this chapter we propose an extension of the WQS where two indices, one looking in the positive and the second in the negative direction, will be built in the same model both at the weights and at the final model estimation step. An application of the new method on the National Health and Nutrition Examination Survey (NHANES) (2015-2016) dietary data and the effect on obesity will be shown besides a simulation study to test the accuracy of the method.

4.1 Model and Methods

The WQS regression, which general formula is shown in equation 2.1, requires that data are split in a training and validation dataset. The first part of the data is used for the weights estimate that allow to build the WQS index while on the remaining part is fitted the final model to test the effect of the score on the outcome. In this study we propose to include two indices in the same model to allow an estimate of the mixture effect both in a positive and a negative direction at the same time both at training and validation step. The new general formula will be the following:

$$g(\mu) = \beta_0 + \beta_{1p}\left(\sum_{i=1}^{c} w_{pi} q_i\right) + \beta_{1n}\left(\sum_{i=1}^{c} w_{ni} q_i\right) + \boldsymbol{z'\varphi} \tag{4.1}$$

where w_{pi} and w_{ni} are the weights associated to each component for the positive and negative direction respectively while β_{pi} and β_{ni} are the two parameters that measure the positive and negative effect of the mixture on the outcome. The two indices will be kept in the model both at the first step where two set of weights are estimated (one for the positive and one for the negative direction) and at the second step when the final model is fitted. The equality and inequality constraints will be applied to both the sets of weights besides a constraint to each β_1 parameter: $\beta_{1p} \geq 0$ and $\beta_{1n} \leq 0$. A penalization term was also introduced to better discriminate between the elements having an effect and those not associated to the outcome and to reduce the noise produced by the null components that can increase the correlation between the two indices. The objective function will be of the form:

$$\begin{aligned} \hat{\theta} = \underset{\theta}{\operatorname{argmin}} \Bigg[\sum_{j=1}^{n} \Bigg(y_j - \Bigg(\beta_0 + \beta_{1p} \sum_{i=1}^{c} w_{pi} q_{pi} + \\ \beta_{1n} \sum_{i=1}^{c} w_{ni} q_{ni} + \boldsymbol{z'\varphi} \Bigg) \Bigg)^2 + \lambda \Bigg(\sum_{i=1}^{c} \mid \nu_{pi} \mid + \sum_{i=1}^{c} \mid \nu_{ni} \mid \Bigg) \Bigg] \end{aligned} \tag{4.2}$$

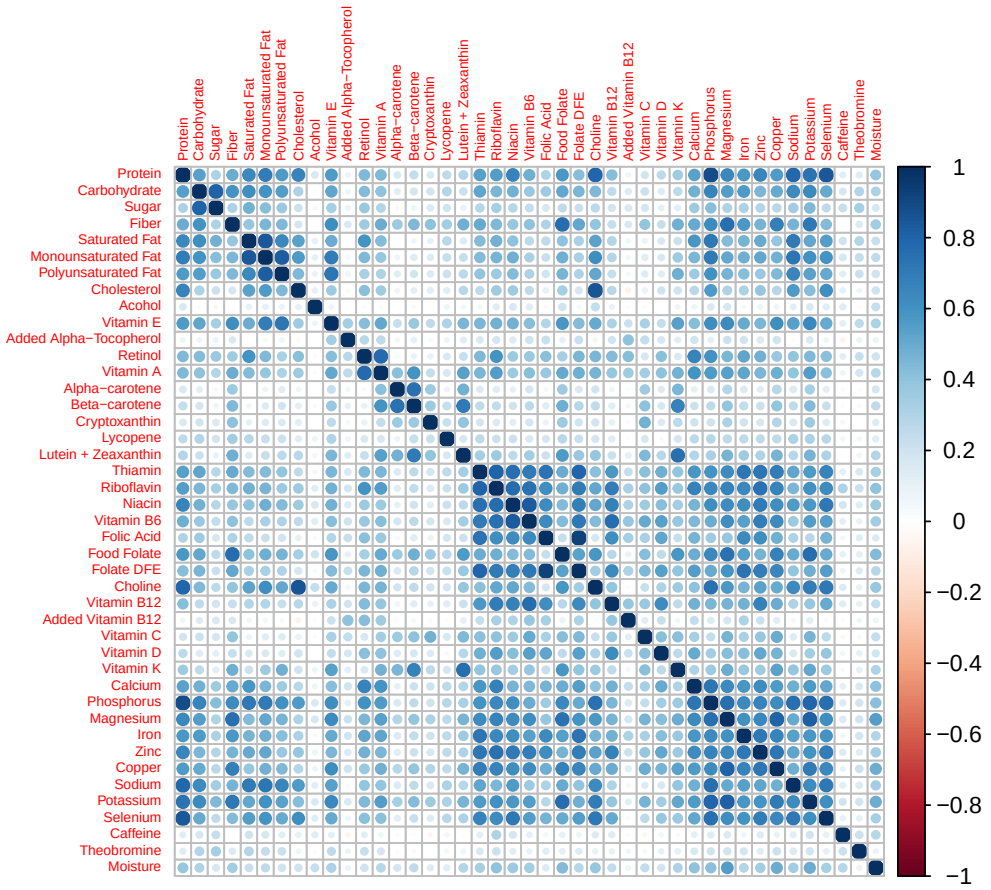
where $\theta = (\beta_0, \beta_{1p}, \beta_{1n}, \nu_{p1}, ..., \nu_{pc}, \nu_{n1}, ..., \nu_{nc}, \varphi, \lambda)$ and $w_i = \frac{|\nu_i|}{\sum_{i=1}^{c} |\nu_i|}$, $\nu_i \in \Re$, $i = 1, ..., c$.
The final model will be: $g(\mu) = \beta_0 + \beta_{1p} WQS_p + \beta_{1n} WQS_n + \boldsymbol{z'\varphi}$.

To further control for the collinearity between the two indices we proposed to apply a different signal function when averaging the weights estimated in each bootstrapped sample. In particular we considered the tolerance (the inverse of the variance inflation factor (VIF)) as the weight in the weighted mean: $f(\hat{\beta}_{1(b)}) = (tol(\hat{\beta}_{1(b)}) / \sum_{b=1}^{B} tol(\hat{\beta}_{1(b)}))^k$ where tol is the tolerance associated to the parameter $\hat{\beta}_{1(b)}$ and k can be chosen depending on the variability of the tolerance values: lower variability would bring to apply higher k to better discriminate between those models where there is a higher collinearity between the positive and the negative index and those where collinearity is less severe.

4.1.1 Simulation Study

In order to show how this new model improves both the estimate of the regression parameters and the weights we performed a simulation study. We took the data from the NHANES 2015-2016 survey cycle where a total of 43 nutrients were measured through the administration of a food frequency questionnaire. In total 100 different datasets were built generating the 43 variables from a multivariate normal distribution keeping the same correlation structure of the original data. As we can see from figure 4.1 the nutrient data show a complex correlation matrix ranging from -0.05 to 0.93.

Figure 4.1: Correlation matrix among the 43 nutrients from the NHANES 2015-2016 survey cycle.



Five nutrients were selected to have weights different from 0 for each direction. In particular the dependent variable was generated from a normal distribution with mean equal to the combination obtained applying the parameters showed in table 4.1 as per WQS with double index formula and a unit standard deviation.

We then fit three different methods on each dataset to compare their performances: in method 1 two separate WQS regressions were built, one exploring the positive and the second the negative direction; in method 2 the WQS set of weights were estimated separately for the positive and negative directions, the two WQS indices were then included in the same regression model fitted on the validation set; we then compared these results with those obtained applying the WQS regression with double index that we called method 3.

Table 4.1: Values of the parameter regression (β_i) and weights (w_i) used to generate the dependent variable. The weight of all the remaining variables not included in the table were set to 0.

Parameter	**PWQS**	**NWQS**
β_{1p}	0.5	
β_{1n}		-0.5
w_1	0	0.05
w_8	0.1	0
w_{10}	0.15	0
w_{19}	0	0.1
w_{21}	0.2	0
w_{27}	0.25	0
w_{34}	0	0.2
w_{35}	0	0.25
w_{40}	0.3	0
w_{41}	0	0.4

4.1.2 Case Study

In the case study we considered the nutrition information from the NHANES 2015-2016 study cycle and we assessed the effects of nutrients on obesity among adults. Obesity was defined as BMI greater or equal to $30kg/m^2$. We excluded subjects with severe obesity ($BMI \geq 40kg/m^2$).
Nutrients were estimated from the dietary intake data that considered the types and amounts of foods and beverages (including all types of water) consumed during the 24-hour period prior to the interview (midnight to midnight). Two interviews were performed: the first one was collected in-person while the second interview was collected by telephone 3 to 10 days later. Details of the survey are described elsewhere (CDC 2016*a*,*b*). In this study we averaged the two nutrition measures and we added the dietary supplement intake when applicable.
Other information was considered in the study like age, sex, race, education as the highest grade or level of school completed or the highest degree received, the ratio of family income to poverty (using the Department of Health and Human Services (HHS) poverty guidelines), the minutes of sedentary activity represented by the time spent sitting on a typical day and the minutes of moderate and vigorous activities spent either during work or during recreational activities categorized using its tertiles (because of the skewed distribution) and the smoking status as never-smokers (subjects who did not smoke as many as 100 cigarettes in their lifetime), former smokers (those who smoked at least 100 cigarettes in their lifetime but were not currently smoking

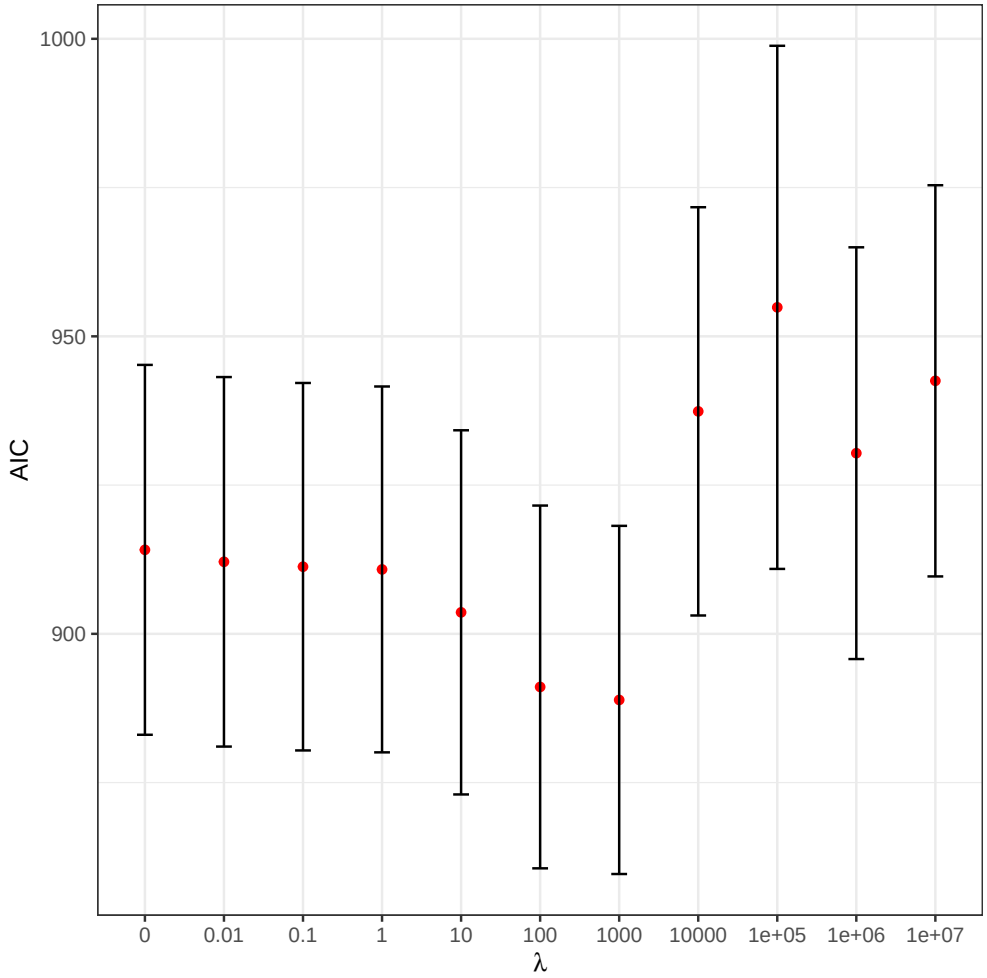
cigarettes), and current smokers (subjects that currently smoked cigarettes). Exclusion criteria were being on any kind of diet to lose weight or for another health-related reason at the time of the interview, having a BMI greater than $40 kg/m^2$ and to be younger than 20 or older than 60 years old. In total 1851 subjects were included in the study.
Kruskal-Wallis test and Chi-squared test were used to test differences between obese and non-obese participants for continuous and categorical variables respectively.

4.2 Results

4.2.1 Simulation Study

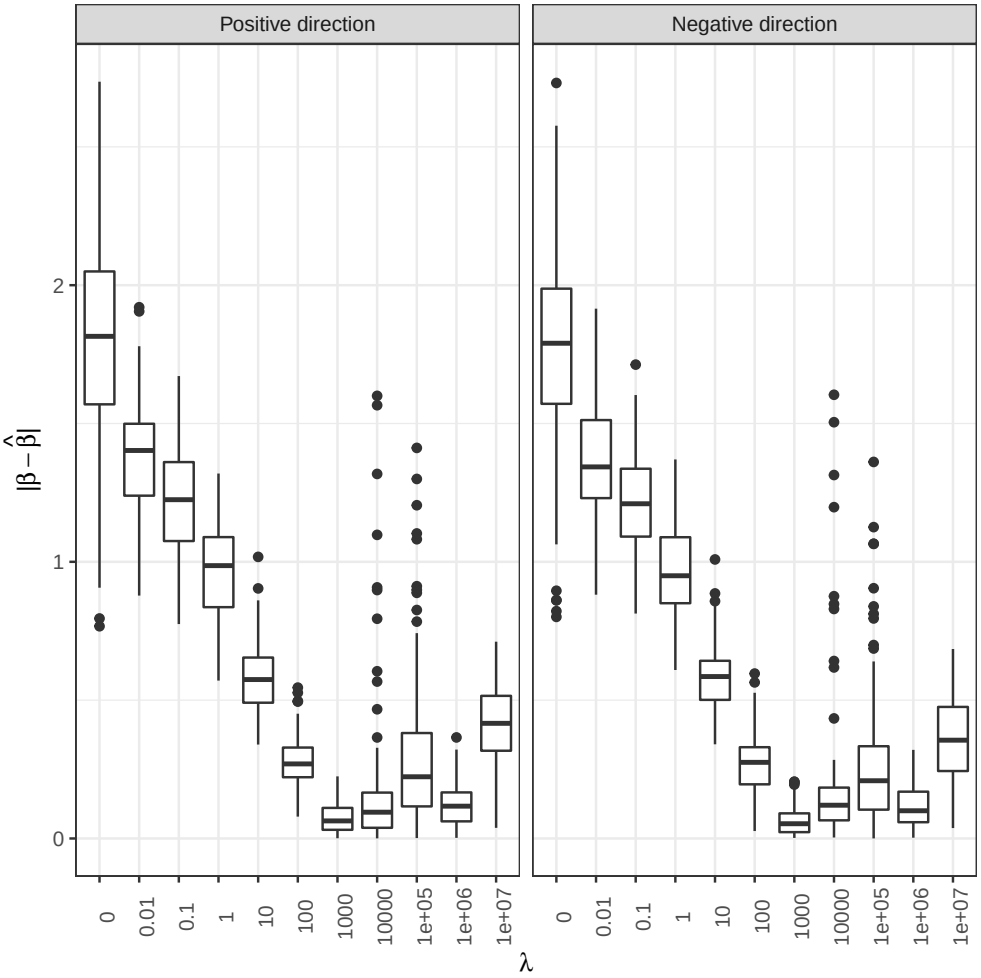
As a first step we tested for the best shrinkage parameter λ: a WQS regression was fitted on each dataset letting λ varying among the values $0, 0.01, 0.1, ..., 10^7$. To choose the best shrinkage parameter we looked at the Akaike Information Criterion (AIC) and we selected the parameter that minimizes it. In our case $\lambda = 1000$ was the optimum value as shown by figure 4.2.

Figure 4.2: WQS regression AIC depending by the shrinkage parameter λ. The red dots represent the average AIC obtained by fitting WQS models on the 100 datasets fixing λ to the corresponding value.



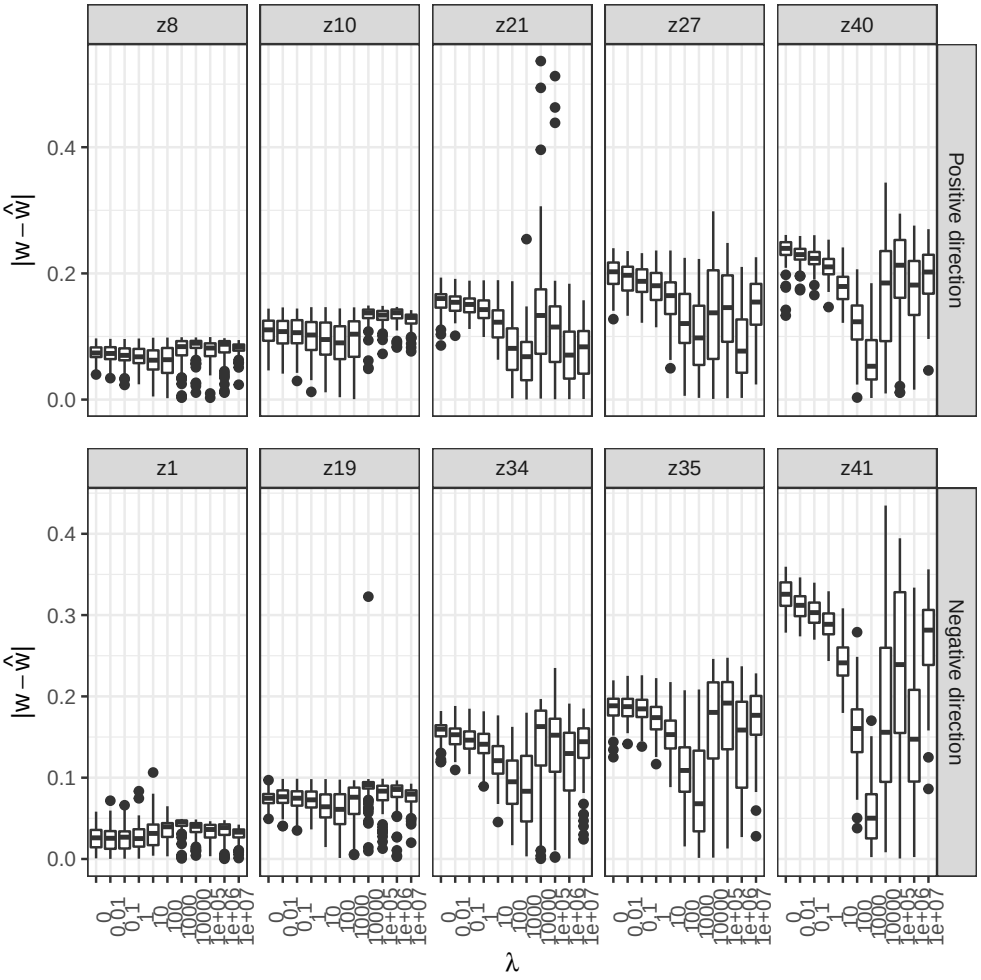
We then checked the accuracy in the parameter estimates at varying shrinkage parameter values. In figure 4.3 the absolute difference between the true value of the regression parameters β_{1p} and β_{1n} for the positive and the negative direction respectively at a fixed λ. The most accurate estimates for the regression parameters correspond to the shrinkage parameter that minimized the AIC ($\lambda = 1000$).

Figure 4.3: Box-plots of the absolute difference between the true value and estimates of the β_{1p} and β_{1n} regression parameters for the positive and negative direction respectively at different shrinkage parameters λ.



We performed the same analysis for the weights: the weight estimates showed the lowest median error the 72.1% of the time (31 times out of 43) when $\lambda = 1000$. In figure 4.4 only the elements set with a weight greater than 0 are shown since these were the situations where a more evident difference was shown among the different λ.

Figure 4.4: Box-plots of the absolute difference between the weight estimates and the true value at varying shrinkage parameter λ. Only the elements that were set with a weight different from 0 are shown.



Once the optimum shrinkage parameter λ was identified, method 1, 2 and 3 were fitted on each of the 100 simulated datasets. Figure 4.5 shows the box-plots of the absolute value of the difference between the true β value and the estimates of the three different models for the positive and negative directions. We can see how the WQS regression with the double index is able to give a better estimate of the parameters compared to method 1 and 2: for positive direction method 3 reduced the mean error of the 62.1% and the 29.5% while for negative direction it reduced the mean error of the 52.9% and the 23.4% compared to method 1 and 2 respectively.

Figure 4.5: Distribution of the absolute difference between the true values of the regression parameter and the estimated value by the three models: method 1 fits two separate WQS regression, one exploring the positive direction and the second exploring the negative direction; method 2 estimates the WQS set of weights separately for the positive and negative directions and then includes the two WQS indices in the same regression; method 3 considers a double index in both the training and validation steps.

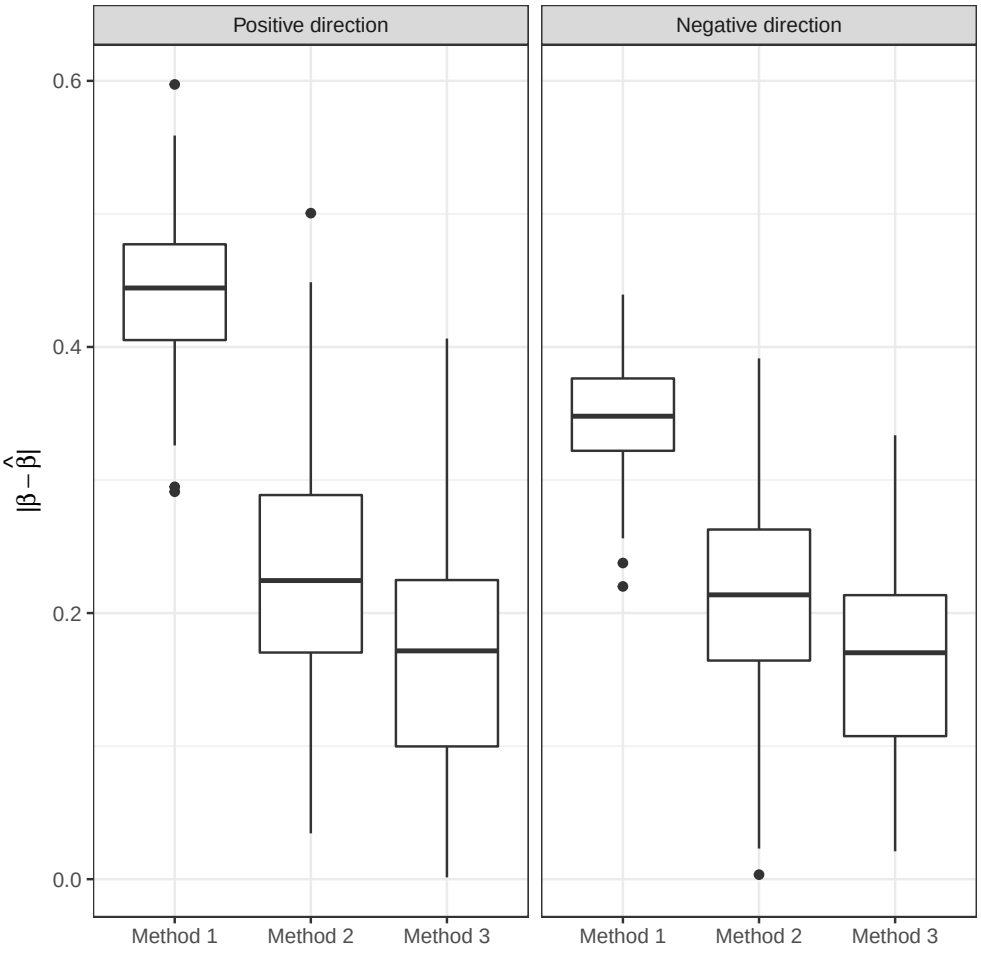
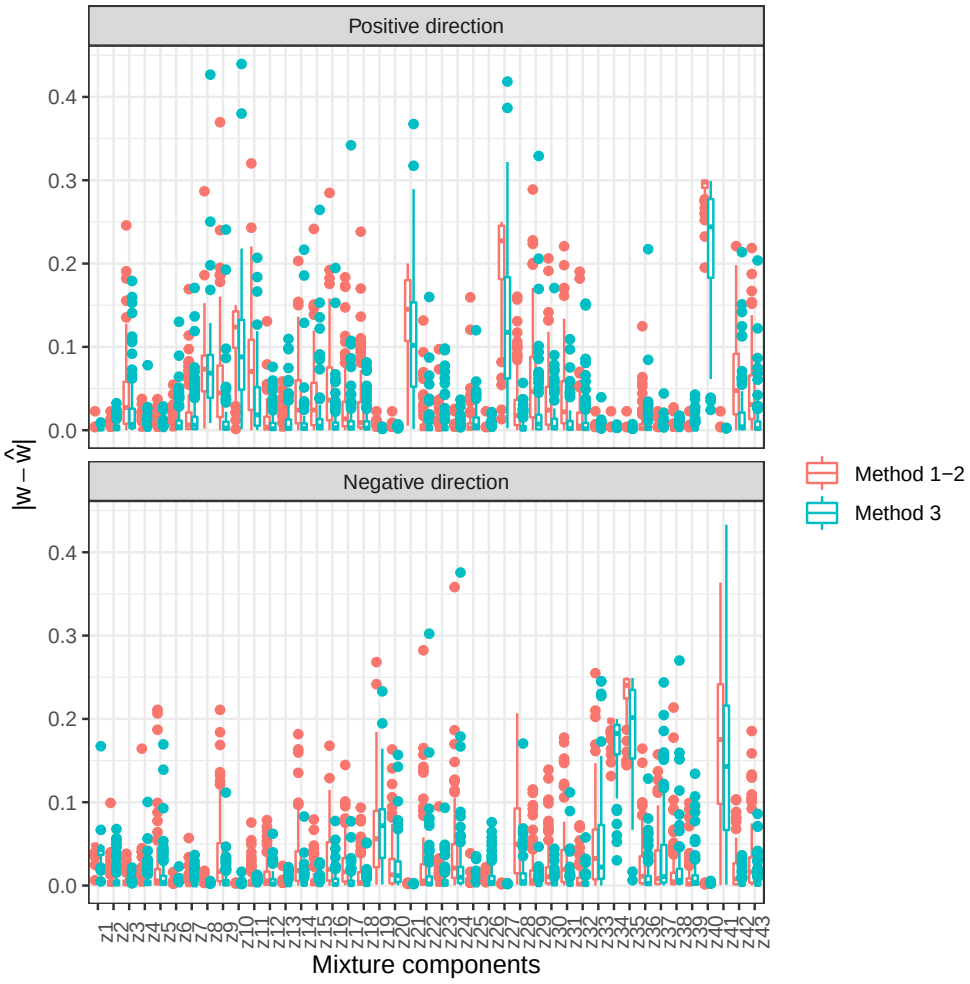


Figure 4.6 shows the performance in measuring the weights of method 3 and method 1 and 2 (that share the same weight estimates). The absolute value of the difference between the true weight and the estimated values was considered. Also in this case we can appreciate how method 3 performs better than the other two regressions showing more accurate weight estimates: in general method 3 shows a lower bias and when we consider the elements with a weight greater than 0 we can observe a reduction of the mean error of 12.5%, 24.3%, 35.7%, 18.5%, 22.6%, 12.9%, 18.2%, and 11.1% for z_{41}, z_{40}, z_{27}, z_{35}, z_{21}, z_{34}, z_{10} and z_1 respectively while there was an increase of the mean error of the 4.3% and 7.6% for z_8 and z_{19} respectively.

Figure 4.6: Distribution of the absolute difference between the true values of the WQS weights and the estimated value by method 1-2 (that share the same weights) and method 3.



4.2.2 Case Study

We then applied the new method of the WQS regression with double index in a real case study. Table 4.2 shows the descriptive statistics of the overall population and divided by obese and non-obese for the covariates included in the WQS regression with double index. A total of 638 (34.5%) subjects were obese and were characterized by a higher median age and higher prevalence of females compared to non-obese participants. A different race distribution was also observed showing a higher percentage of Mexican and Black subjects and a lower prevalence of Asian participants among obese. Finally, a lower level of education was detected in obese people.

Table 4.2: Descriptive statistics of the variables included in the study for the overall population and divided by obese and non-obese. Median, 1^{st} and 3^{rd} quartiles are shown for continuous variables while counts and percentages were considered for categorical variables. Kruskal-Wallis test and Chi-squared test were used to test differences for continuous and categorical variables respectively.

	Non-Obese (N=1213)	**Obese (N=638)**	**Total (N=1851)**	**p-value**
Age	38.0 (28.0, 49.0)	42.0 (32.0, 51.0)	40.0 (30.0, 50.0)	<0.001
Sex				<0.001
M	631 (52.0%)	285 (44.7%)	916 (49.5%)	
F	582 (48.0%)	353 (55.3%)	935 (50.5%)	
Race				<0.001
Mexican	164 (13.5%)	133 (20.8%)	297 (16.0%)	
Other Hispanic	138 (11.4%)	82 (12.9%)	220 (11.9%)	
White	416 (34.3%)	198 (31.0%)	614 (33.2%)	
Black	236 (19.5%)	165 (25.9%)	401 (21.7%)	
Asian	220 (18.1%)	27 (4.2%)	247 (13.3%)	
Others	39 (3.2%)	33 (5.2%)	72 (3.9%)	
Education	4.0 (3.0, 5.0)	4.0 (3.0, 4.0)	4.0 (3.0, 5.0)	<0.001
Family income to poverty ratio	2.2 (1.2, 4.2)	2.1 (1.1, 3.8)	2.2 (1.2, 4.1)	0.072
Minutes of sedentary activity	360.0 (180.0, 480.0)	360.0 (240.0, 480.0)	360.0 (240.0, 480.0)	0.253
Moderate and Vigorous-Intensity Activities				0.352
Low	396 (32.6%)	229 (35.9%)	625 (33.8%)	
Medium	436 (35.9%)	214 (33.5%)	650 (35.1%)	
High	381 (31.4%)	195 (30.6%)	576 (31.1%)	
Smoking status				0.345
Never	754 (62.2%)	387 (60.7%)	1141 (61.6%)	
Former	187 (15.4%)	115 (18.0%)	302 (16.3%)	
Current	272 (22.4%)	136 (21.3%)	408 (22.0%)	

In table 4.3 are shown the summary statistics related to the 43 nutrients included in the analysis. All the elements that showed a significant difference between the two groups had higher values among non-obese subjects.

Table 4.3: Summary statistics of the 43 nutrients included in the analysis. Median, 1^{st} and 3^{rd} quartiles are shown for the overall population and divided in obese and non-obese participants. Kruskal-Wallis test was applied to test for differences between the two groups.

	Non-Obese (N=1213)	**Obese (N=638)**	**Total (N=1851)**	**p-value**
Added vitamin B12 (mcg)	0.0 (0.0, 1.3)	0.0 (0.0, 0.9)	0.0 (0.0, 1.2)	0.171
Added Vitamin E (alpha-tocopherol) (mg)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.069
Alcohol (gm)	0.0 (0.0, 9.3)	0.0 (0.0, 6.7)	0.0 (0.0, 8.5)	**0.024**
Alpha-carotene (mcg)	83.5 (22.5, 501.0)	55.0 (17.0, 291.0)	72.5 (21.0, 432.8)	**<0.001**
Beta-carotene (mcg)	1118.0 (479.0, 2973.5)	804.2 (350.2, 1897.5)	992.0 (417.8, 2528.5)	**<0.001**
Caffeine (mg)	85.0 (27.5, 176.0)	91.8 (24.4, 193.0)	88.0 (26.8, 182.0)	0.248

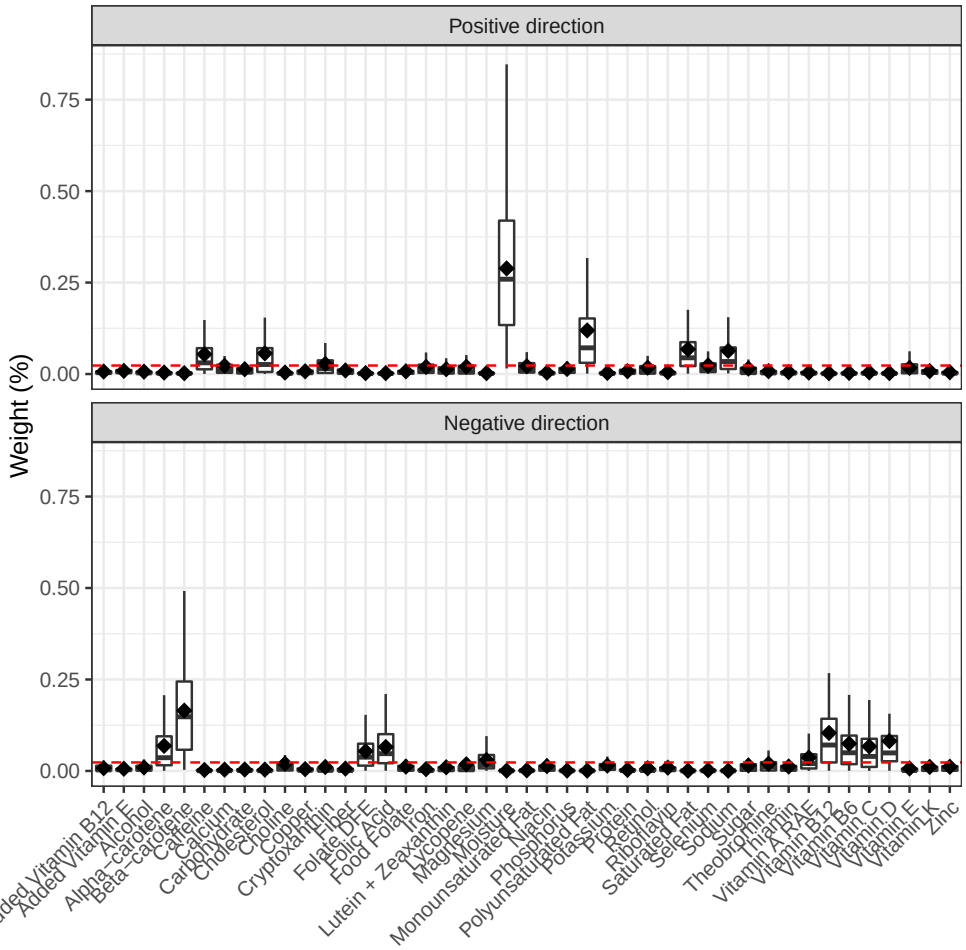
Calcium (mg)	911.0 (639.0, 1264.0)	883.2 (633.0, 1229.9)	902.5 (638.2, 1257.0)	0.290
Carbohydrate (gm)	244.3 (181.7, 312.2)	235.0 (173.8, 298.1)	241.4 (178.6, 307.6)	**0.043**
Cholesterol (mg)	265.0 (169.5, 389.5)	267.4 (169.8, 401.8)	265.5 (169.5, 392.8)	0.597
Total choline (mg)	314.7 (227.0, 414.6)	296.8 (212.4, 410.2)	306.9 (221.8, 413.1)	**0.027**
Copper (mg)	1.2 (0.9, 1.7)	1.1 (0.8, 1.6)	1.2 (0.8, 1.7)	**<0.001**
Beta-cryptoxanthin (mcg)	43.0 (16.5, 104.0)	46.8 (15.5, 92.4)	44.0 (16.0, 99.5)	0.828
Dietary fiber (gm)	15.9 (11.0, 22.6)	15.1 (10.4, 20.9)	15.7 (10.8, 22.1)	**0.026**
Folate, DFE (mcg)	587.0 (377.5, 878.8)	502.0 (325.0, 759.9)	547.2 (358.8, 843.0)	**<0.001**
Folic acid (mcg)	195.7 (106.5, 364.0)	157.8 (92.5, 294.9)	182.5 (99.2, 343.5)	**<0.001**
Food folate (mcg)	202.0 (138.0, 289.5)	184.5 (131.5, 268.5)	196.0 (135.8, 283.0)	**0.004**
Iron (mg)	14.1 (10.1, 19.3)	13.1 (9.6, 18.3)	13.7 (9.9, 19.1)	**0.017**
Lutein + zeaxanthin (mcg)	887.0 (480.5, 1754.0)	769.5 (430.5, 1440.0)	853.0 (464.8, 1644.2)	**0.002**
Lycopene (mcg)	2764.0 (915.0, 6688.0)	2646.2 (878.6, 6613.1)	2718.0 (904.2, 6668.8)	0.495
Magnesium (mg)	297.5 (222.5, 401.0)	274.5 (202.1, 362.0)	291.0 (216.8, 389.0)	**<0.001**
Moisture (gm)	2601.1 (1987.7, 3486.9)	2682.3 (1921.5, 3677.9)	2632.7 (1964.4, 3565.8)	0.459
Monounsaturated fatty acids (gm)	26.7 (18.8, 35.0)	26.2 (18.6, 35.6)	26.5 (18.7, 35.3)	0.959
Niacin (mg)	27.6 (19.8, 38.7)	25.1 (18.2, 35.7)	26.9 (19.2, 37.8)	**0.002**
Phosphorus (mg)	1306.0 (1004.5, 1673.0)	1265.8 (948.5, 1648.4)	1290.5 (983.0, 1670.0)	0.108
Polyunsaturated fatty acids (gm)	16.9 (12.0, 23.6)	17.2 (12.1, 24.7)	17.0 (12.0, 23.9)	0.370
Potassium (mg)	2490.0 (1897.0, 3149.5)	2325.0 (1727.9, 2931.4)	2444.0 (1841.8, 3078.8)	**0.002**
Protein (gm)	78.9 (59.8, 101.6)	75.3 (57.3, 97.9)	77.7 (59.2, 100.0)	**0.041**
Retinol (mcg)	316.5 (188.5, 505.5)	301.8 (172.9, 477.8)	311.5 (183.2, 493.0)	0.180
Riboflavin (Vitamin B2) (mg)	2.1 (1.5, 3.1)	1.9 (1.4, 2.8)	2.0 (1.4, 3.0)	**0.005**
Saturated fatty acids (gm)	23.9 (16.5, 33.3)	24.9 (17.0, 34.0)	24.2 (16.7, 33.5)	0.317
Selenium (mcg)	120.0 (86.0, 163.1)	112.2 (82.7, 153.1)	117.6 (84.3, 159.3)	**0.043**
Sodium (mg)	3409.5 (2458.5, 4338.0)	3336.0 (2518.2, 4170.6)	3377.5 (2470.2, 4280.2)	0.378
Sugars (gm)	95.0 (63.4, 134.6)	93.9 (61.6, 137.0)	94.6 (62.3, 135.3)	0.654
Theobromine (mg)	8.5 (0.0, 39.5)	7.5 (0.0, 35.9)	8.0 (0.0, 38.0)	0.279
Thiamin (Vitamin B1) (mg)	1.7 (1.2, 2.5)	1.5 (1.1, 2.3)	1.6 (1.2, 2.4)	**<0.001**
Vitamin A, RAE (mcg)	501.0 (305.0, 767.0)	432.5 (275.4, 677.0)	477.5 (292.5, 732.8)	**<0.001**
Vitamin B12 (mcg)	5.3 (3.0, 10.6)	4.5 (2.6, 9.0)	5.0 (2.9, 10.1)	**0.001**
Vitamin B6 (mg)	2.2 (1.5, 3.6)	1.9 (1.3, 3.2)	2.1 (1.5, 3.4)	**<0.001**
Vitamin C (mg)	84.5 (36.4, 158.0)	71.2 (30.7, 134.7)	80.6 (35.0, 149.8)	**0.002**
Vitamin D (D2 + D3) (mcg)	5.4 (2.3, 15.1)	4.1 (2.0, 10.5)	4.8 (2.2, 13.4)	**<0.001**
Vitamin E as alpha-tocopherol (mg)	7.8 (5.3, 10.8)	7.3 (5.1, 10.4)	7.6 (5.2, 10.7)	0.083
Vitamin K (mcg)	86.9 (50.8, 148.7)	75.5 (47.6, 124.7)	82.6 (49.2, 140.1)	**<0.001**
Zinc (mg)	11.6 (8.3, 16.8)	10.8 (7.4, 15.6)	11.3 (7.8, 16.3)	**0.002**

We then applied the WQS regression with double index to test for the association between the nutrients and the outcome adjusting for all the covariates reported in table 4.2. We used a repeated holdout approach to have more stable results including all the observations in the study to estimate both the weights and the regression parameters during the repeated testing and validation steps (Tanner et al. 2019). A total of 100 repeated holdout WQS with double index were performed. Table 4 shows the effects and their 95% CIs of both the positive and the negative index on the probability of being obese. Both indices were associated with the outcome. The median was used as the parameter point estimates while the 2.5^{th} and the 97.5^{th} percentiles were considered to build the 95% CIs. In table 4.4 are also shown the medians of the weights greater than the prespecified cutoff (0.023=1/43) for the positive and negative index where moisture, polyunsaturated fatty acid, saturated fatty acid, sodium, caffeine and cholesterol showed a predominant role in the positive direction while beta-carotene, vitamin B12, vitamin B6, vitamin D, folic acid, vitamin C, folate DFE and alpha carotene were inversely associated with obesity. Figure 4.7 represents the mean and the distribution of all the elements included in the analysis estimated for both indices.

Table 4.4: WQS regression with double index results. The estimates and 95% confidence intervals (CI) of the positive (pwqs) and negative (nwqs) indices are shown. Model was adjusted for age, sex, race, education, the ratio of family income to poverty, the minutes of sedentary activity, the minutes of moderate and vigorous activities and the smoking status. The second part of the table shows the magnitude of the weights greater than the prespecified cutoff (0.023) for both the positive and the negative index.

	Estimate	**95% CI**
pwqs	0.084	0.009; 0.198
nwqs	-0.137	-0.243; -0.068
	Weights	
pwqs		
Moisture	0.259	
Polyunsaturated fatty acids	0.072	
Saturated fatty acids	0.044	
Sodium	0.034	
Caffeine	0.031	
Cholesterol	0.026	
nwqs		
Beta-carotene	0.148	
Vitamin B12	0.071	
Vitamin B6	0.050	
Vitamin D	0.049	
Folic acid	0.047	
Vitamin C	0.040	
Folate DFE	0.038	
Alpha-carotene	0.036	

Figure 4.7: Box-plot of the weights associated to the positive and negative index estimate through the repeated holdout WQS regression with double index. The dashed red line represents the prespecified cut-off established to identify the most important elements of the mixture and set equal to the inverse of the number of the mixture components (0.023).



Chapter 5

Application of WQS regression to genetic data

As a final work we tested the performance of WQS in a different context like genetic compared to environmental exposures. Since the advent of DNA hybridization microarrays, a persistent challenge in this field is to analyze and interpret genes or pattern of expression affecting a particular phenotype (e.g. tumor vs normal). This information provides a better understanding of the underlying biological process or can be used to predict the condition on a new sample. Different methods like single sample Gene Set Enrichment Analysis (ssGSEA) (Barbie et al. 2009), Gene Set Variation Analysis (GSVA) (Hanzelmann et al. 2013), Pathway Level Analysis of Gene Expression (PLAGE) (Tomfohr et al. 2005) combining z-scores (Lee et al. 2008) and singscore (Foroutan et al. 2018) have been developed to score individual samples against gene sets. In particular we will consider the ssGSEA method which is the most widely used approach in this context. This method summarizes the gene expression in a single enrichment score for each pairing of sample and gene set independently by the condition. ssGSEA scores a gene set enrichment profile representing the activity level of the biological mechanism where the genes can be up- or down-regulated. The effect of the score on the considered dependent variable can be tested in a standard regression model. This allows to overcome the problem of the high dimensionality of the data that classical statistical methods are not able to manage. However, a limit of all these approaches included ssGSEA is the unsupervised estimate of the score, where by unsupervised we mean that the index is created in absence of the dependent variable. WQS regression is a novel statistical method that can deal with high dimensional data especially through its random sub-

set extension, and it is able to estimate a score in the presence of the outcome better identifying the true genes which expression affects the dependent variable. Moreover, the initial application of the WQS regression with double index allows to identify either the bidirectional or unidirectional effect of the gene signature due to the up- and down-regulated gene set's members.
In this chapter we propose the application of the WQS regression within the context of the biological pathways. We will test its performance compared to the ssGSEA method and we will show how to apply it in a real case study.

5.1 Model and Methods

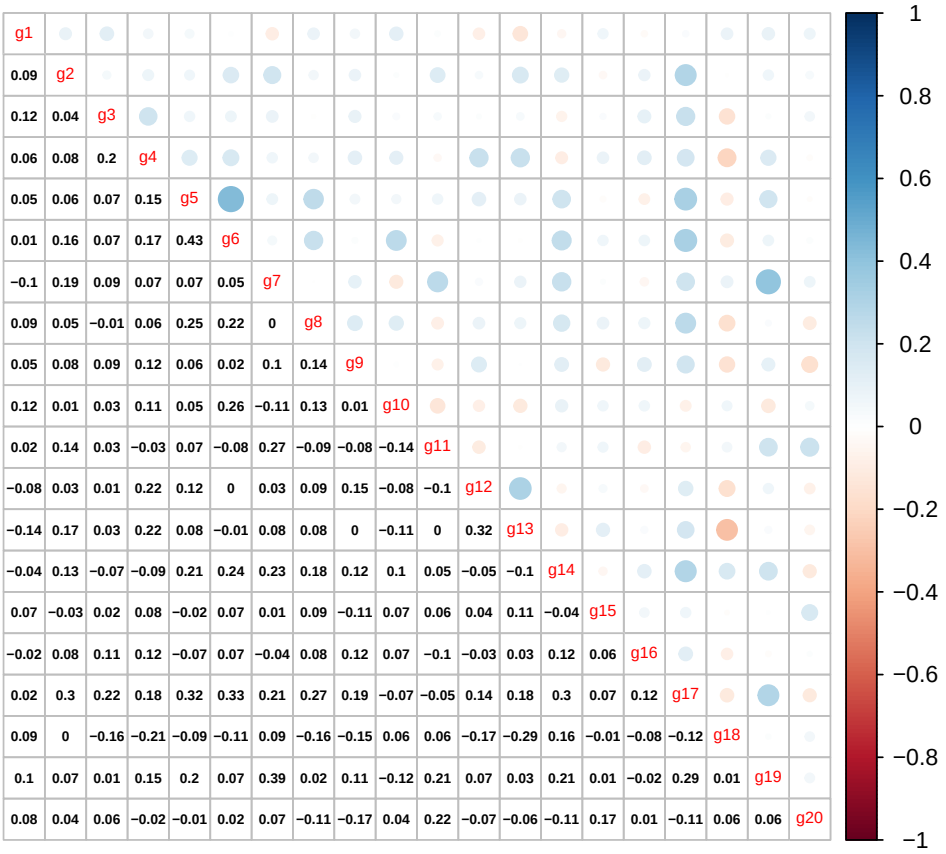
In order to apply the WQS regression in the context of a pathway analysis we propose to apply the WQS with double index as per equation 4.1. In this case the two indices are necessary to identify an effect of the up- or down-regulation of the genes on the considered outcome. A single score may be necessary to represent the pathway hypothesized to affect the condition of interest. In this case a second model has to be fitted. If only one of the two scores is statistically significant then a WQS regression with single index (equation 2.1) has to be performed choosing the same direction of the significant association. If there is an effect of both indices, the genes with the highest weights (selected through the prespecified cutoff τ usually set equal to the inverse of the number of elements in the mixture) associated to the index representing the opposite direction to the one identified by the pathway are multiplied by -1 to reverse their effect. A WQS regression with single index can now be fitted to estimate the association between the gene set and the outcome.

5.1.1 Simulation Study

To test the performance of the WQS in estimating the effect of a chosen gene set on an outcome we considered data from the curatedOvarianData (Ganzfried et al. 2013) from The Cancer Genome Atlas (TCGA) program. This dataset included patients with ovarian cancer providing uniformly prepared microarray data for 2970 patients from 23 studies with documented clinical information. A total of 578 subjects being part of the TCGA study were considered in our analysis. A genetic pathway was randomly selected

and the correlation matrix (figure 5.1) of the variables representing the gene expression was applied to derive the covariance matrix used to define the simulated independent variables from a multivariate normal distribution (with a null mean vector). A total of 20 genes were included in the analysis and 500 observations were generated for each of the 100 datasets.

Figure 5.1: Gene set correlation matrix.



Two outcomes were generated from a normal distribution with a mean equal to the resulting WQS formula setting the parameters as described in table 5.1 and a unit variance. Those weights associated to the genes which did not have an effect on the outcome were not reported in table 5.1 and were set to zero. The first dependent variable y_1 was built under the assumption of a double directionality of the association while y_2 was identified considering only a positive effect of the gene set. In total 8 genes were selected for y_1, 4 in each direction. The same 4 genes included in the positive index for y_1

were considered for y_2.

Table 5.1: Parameter values used to generate the two dependent variables y_1 and y_2 in the presence of a double directionality of the association and only a positive effect of the gene set respectively.

	y_1	y_2
β_{1p}	0.5	0.5
β_{1n}	0.5	0
W_{3p}	0.15	0.15
W_{14p}	0.5	0.5
W_{15p}	0.4	0.4
W_{19p}	0.25	0.25
W_{2n}	0.3	0
W_{7n}	0.15	0
W_{11n}	0.45	0
W_{12n}	0.1	0

5.1.2 Case Study

We then applied the ssGSEA and the WQS regression on the real data from the curatedOvarianData. We selected a known pathway that increases the risk of death among patients with ovarian cancer. The pathway involved in cell-cycle is known to be altered at high frequencies in ovarian cancer among others: it identifies a series of events that control the genome replication and the subsequent chromosomes segregation into daughter cells (Matthews 2011). A total of 527 genes were identified as involved in the cell-cycle among those present in the curatedOvarianData; as previously mentioned the sample size was made of 578 observations which reduced to 559 subjects because of few missing values of the dependent variable. Because of the large number of variables (genes) included in the analysis we applied a WQS_{RS} while a WQS_{RH} was considered to allow all observations to be chosen in both training and validation steps. The results obtained from applying the ssGSEA gene score in a logistic regression and the WQS regression to test for the effect of the expression of the selected gene set and the condition of each subject (deceased or living) were then compared.

5.2 Results

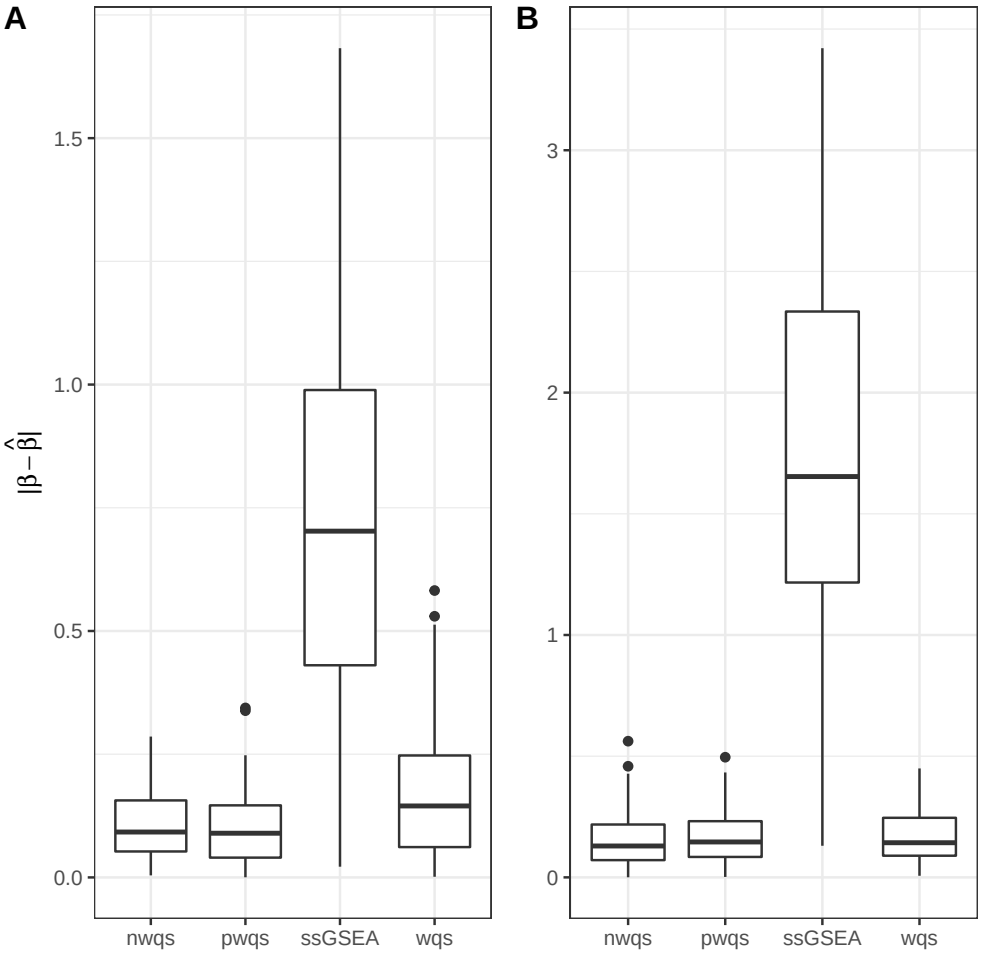
5.2.1 Simulation Study

In the first instance we applied the WQS regression with double index and the ssGSEA to the simulated data considering y_1 as the outcome. In figure 5.2A

are represented the boxplots of the absolute differences between the estimates of the regression parameter associated to the score for each method and the true values. The WQS with double index shows more accurate estimates of the regression parameters (ME: 0.100; SD: 0.074 and ME: 0.105; SD: 0.069 for pwqs and nwqs scores respectively) compared to ssGSEA (ME: 0.721; SD: 0.406) showing a bias reduction of the 85.4%. Once we reverse the variables associated to the down-regulated genes with a weight greater than the prespecified cutoff we can still see a better performance of the WQS with single index (ME: 0.164; SD: 0.125) reducing the bias of the 77.3%. The effects of the ssGSEA and the WQS with single index were compared to 1 since the total effect (the sum of the absolute value of both positive and negative directions) sums to the unit value. Moreover, in the presence of a double effect of the up- and down-regulated selected genes we can see an underestimated effect (average effect: 0.299) and higher SEs (average SE: 0.440) when applying the ssGSEA compared to the WQS with single index where the mean effect is slightly lower (0.913) with an associated lower SEs (average SE: 0.188).

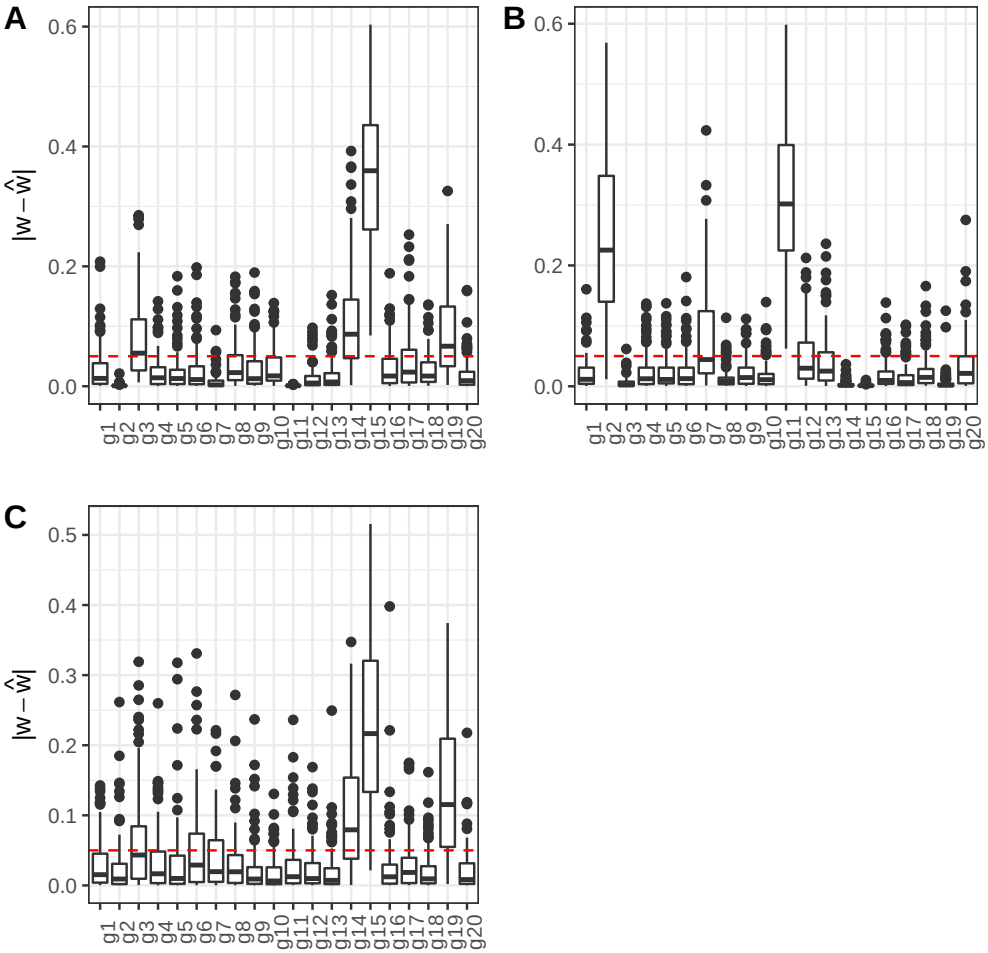
In figure 5.2B we compared the results of the two methods applied to the outcome y_2 where only a positive effect is simulated. Also in this case we can observe more accurate estimates of the WQS with double index (ME: 0.169, SD: 0.113 and ME: 0.148; SD: 0.111) compared to ssGSEA (ME: 1.755, SD: 0.768) showing a reduced bias of the 90.4%. The WQS with double index did not detect a negative effect of the score on the dependent variable; we then applied the WQS with single index setting a positive direction (ME: 0.172; SD: 0.109) and we could still see a better performance compared to ssGSEA reducing the mean error of the 90.2%. In the situation of an effect in only one direction (due to either the up- or down-regulated genes) we observed an overestimated effect of the ssGSEA (average effect: 2.2453) and higher SEs (average SE: 0.791) compared to WQS with single index where we can see a slightly higher mean effect (average effect: 0.516) and lower SEs (average SE: 0.204).

Figure 5.2: Boxplots of the absolute difference between the true and estimated value of the regression parameter associated to the score for each method (WQS with double index (pWQS and nWQS), ssGSEA and WQS with single index) applied to test the effect on the two outcomes y_1 and y_2.



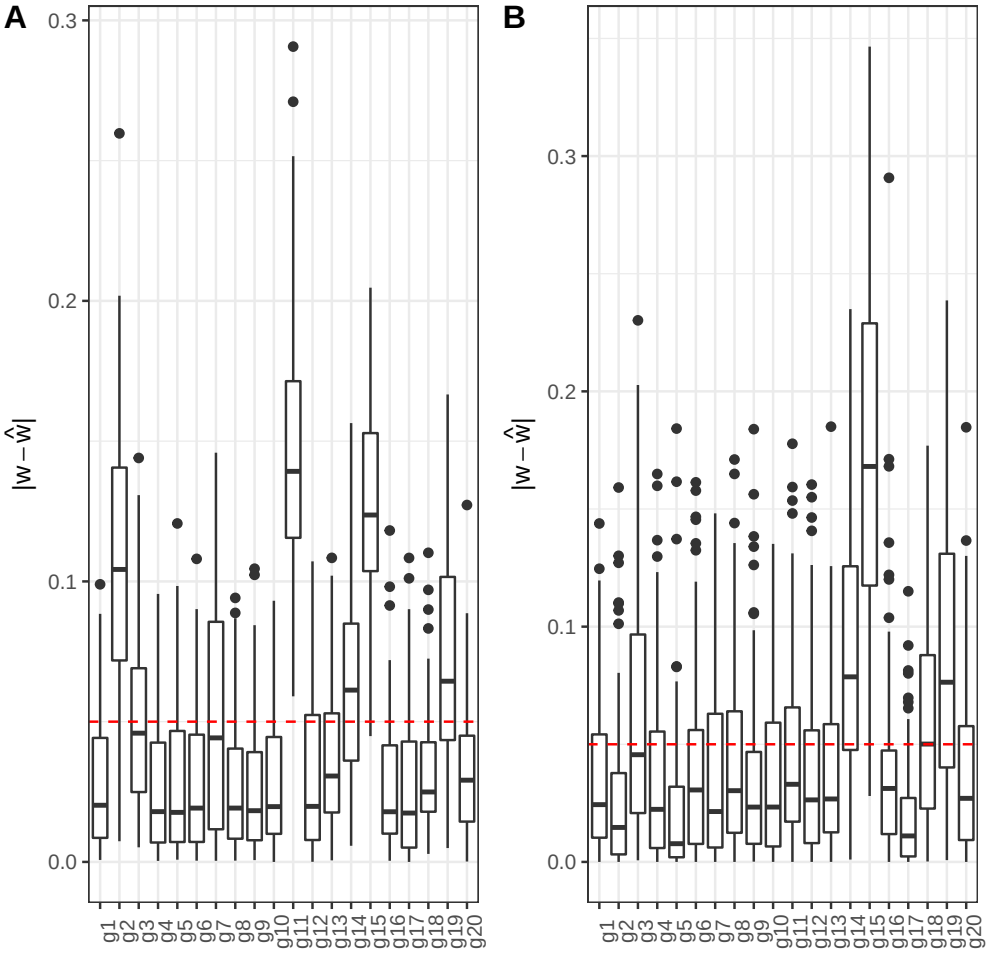
Through WQS regression we are also able to identify the genes that show a greater effect in the identified pathway. As per table 5.1 g_{15}, g_{19}, g_{14} and g_3 were identified as the genes with a positive significant effect on the dependent variable and were correctly selected 100%, 62%, 72% and 56% of the time respectively by WQS regression. The WQS negative score was then built giving a non-null weight to g_{11}, g_2, g_7 and g_{12} which were correctly chosen 100%, 98%, 49% and 35% of the time respectively. The results related to the outcome y_2 showed that g_{15}, g_{19}, g_{14} and g_3 were correctly identified the 93%, 77%, 69% and 45% of the time. The distribution of the weights of the three scores is shown in figure 5.3A-B-C respectively.

Figure 5.3: Boxplots of the absolute difference between the true and estimated value of the weights from the WQS regression with double index applied to the outcome y_1 (panel A for positive score and panel B for negative score) and y_2 (panel C for positive score; negative score weights are not shown since the effect in this direction was not significant).



Once the distribution of the down-regulated genes was reversed to estimate the effect of a single score on y_1 we observed that the genes g_{11}, g_{15}, g_2, g_{19}, g_{14}, g_3, g_7 and g_{12} were correctly selected the 100%, 99%, 91%, 69%, 65%, 46%, 49% and 28% of the time respectively. The results related to the outcome y_2 showed that g_{15}, g_{19}, g_{14} and g_3 were correctly identified the 97%, 68%, 72% and 49% of the time. The distribution of the weights of the two scores is shown in figure 5.4A-B respectively.

Figure 5.4: Boxplots of the absolute difference between the true and estimated value of the weights from the WQS regression with single index applied to the outcome y_1 (panel A) and y_2 (panel B).



5.2.2 Case Study

The study moved to the application of the WQS with double index on a real case study: a total of 527 genes being part of the cell-cycle pathway were selected to test the association between the gene set and the condition of each subject (deceased or living). When we applied the ssGSEA to estimate the genetic score and test its effect on the outcome through a logistic regression we find a negative but not statistically significant effect (β -1.017; 95%CI -2.356, 0.300). The WQS regression with double index showed a significant association in the negative but not in the positive direction (β -0.230; 95% CI -0.399, -0.055 and β 0.157; 95% CI -0.111, 0.416 respectively) (results displayed in table 5.2).When we applied the WQS regression with single index after reversing the distribution of the down-regulated genes we observed an

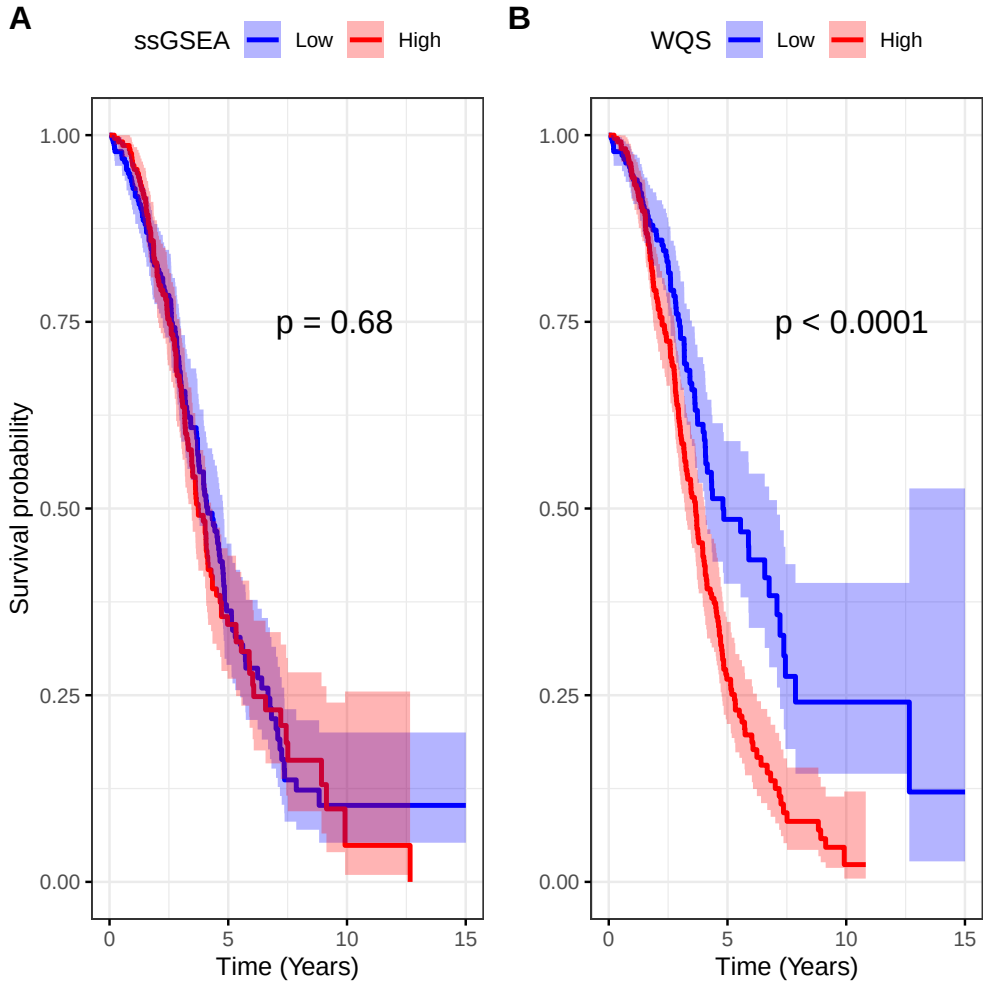
increased risk of death corresponding to higher score values (β 0.923; 95% CI 0.499, 1.397).

Table 5.2: Estimates and 95% CI obtained applying the ssGSEA, the WQS with double index and the WQS with single index (built after reversing the down-regulated genes).

	Estimates	95% CI
ssGSEA	-1.017	-2.356, 0.300
pwqs	0.157	-0.111, 0.416
nwqs	-0.230	-0.399, -0.055
wqs	0.923	0.499, 1.397

In order to represent the difference in detecting the effect of the gene expression on the outcome we categorized the ssGSEA and WQS in "Low" and "High" score using the respective medians and then looked at the effect on time to death. Comparing the survival probability between the two groups we can clearly see how those subjects with a high WQS score have a lower probability to survive than those with a lower score ($p<0.0001$) while we do not see a significant difference when using the ssGSEA score ($p=0.680$) (figure 5.5).

Figure 5.5: Effect of ssGSEA (A) and WQS (B) score on time to death.



Chapter 6

Discussion

6.1 Effect modification in WQS regression

Through the introduction of an interaction term in the WQS regression we were able to test for the effect modification of an environmental mixture due to a covariate of interest (either categorical or continuous). WQS regression accommodates the evaluation of the mixture effect as a weighted index of multiple elements which may have a complex correlation pattern. Once the weights are estimated using the training data, the index is built and included in analysis of the hold-out validation data as a continuous variable. Through representation of the chemical mixture in a weighted index, we can easily test for the interaction with a second covariate adding a multiplicative term as in a standard regression. The analysis strategy permits a direct test for changes in the mixture effect due to important covariates (e.g., age, sex).
In the simulation study we evaluated the importance of including the interaction term while estimating the weights to obtain a lower bias on both the regression parameter and weight estimates. In the case of a categorical effect modifier we showed that when considering both the interaction between the WQS index and the categorical variable, and we stratify the weights evaluating a set of parameters for each level of the categorical variable, we obtain more accurate estimates. Moreover, when applying method 3 in the absence of an interaction or when weights do not vary across categories we can appreciate similar results to method 1 and 2. Based on these simulations we recommend to initially fit the complex model including both the interaction term and stratified weights and then move to a simpler model if the interaction term is not statistically significant; or, in the case of categorical interaction terms, if the estimated weights are similar across categorical vari-

able levels. In the first case the stratified weights can be kept even if the interaction term is not significant, while in the second situation we can specify the model with only the interaction term between the WQS index and the categorical variable without stratifying the weights.
In the case study, we applied the extension of WQS regression allowing for interaction where we tested for the mixture effect of blood concentrations of Pb, Cd, Hg, Se and Mn on FVC among children from 6 to 17 years old. Using a mixture approach, we saw an inverse association of metal exposure with FVC mainly driven by Hg, Pb and Mn. In the previous work from Madrigal et al. (2018), only a direct effect of Mn on FVC among blood metal concentrations was detected. When including an interaction between Mn and age they also observed that this association was strongest among older youth.
Through the WQS regression modeling, we see how the mixture effect can change compared to a single element analysis: WQS regression allowed us to find that not only Mn but also Hg and Pb have a meaningful effect on FVC when considering all metals in the same analysis - evidence of a mixture effect. Moreover, when we include the interaction with age we were able to find that the association between FVC and Pb and Hg was attenuated among older children. This agrees with previous literature stressing the higher vulnerability of younger children to environmental chemical exposure (Landrigan et al. 2016, Sly et al. 2016, Suk et al. 2016). Finally, we were also able to find different effects of metal exposure between males and females. In particular the model demonstrated a higher effect of the mixture among females mainly driven by Cd and Hg meaning that female FVC is more susceptible to these two metals compared to males.
A previous work from Lee et al. (2019) (Lee et al. 2019) demonstrated another approach to estimating stratum-specific weights of exposures in a mixture: the variables representing the multiplicative interaction between the chemicals and the effect modifiers are included in the WQS index as part of the mixture and a weight is also attributed to these components. A potential disadvantage in such an approach is that it does not allow for a direct hypothesis test for the significance of putative interactive effects. In contrast, the approach proposed here allows for direct tests for effect modification at the level of the whole mixture and the outcome, not at the interaction of the single elements. This is particularly useful when we are interested in considering the impact of different elements as a whole and the starting hypothesis

is that the effect modifier can interact with all the chemicals included in the mixture. Moreover, through the stratified weights we were able to measure the category specific contribution of each element to the effect of the whole mixture when considering a categorical modifier.

6.2 WQS with double index

Through the introduction of a double index we extended the WQS regression to the case where the mixture considered in the study can have both a positive or a negative effect. We allowed to build two scores in the same regression model both including all the elements of the mixture, one constrained to be positive and one to be negative. A recent work from Keil et al. (2020) (Keil et al. 2020) introduced a new approach that allows to estimate the overall effect of the mixture on the outcome when there is uncertainty about the effect direction of some exposures. Differently from the original WQS regression they proposed to estimate positive and negative weights within the same index using normalized linear (or generalized linear) regression coefficients and then estimating the mixture effect via a standard g-computation algorithm. Thanks to this approach we are able to measure the overall effect of the mixture on the outcome but we cannot estimate the impact in the positive and negative direction. Through our method we introduced the possibility to measure both the beneficial and harmful effect of the mixture exposure.
In this work we showed how the two indices were built and how we deal with the correlation between the two scores: because of the high correlation among the elements included in the mixture we noticed an increased risk of collinearity when including both indices in the same regression model. To face this problem we applied two strategies. At first we introduced a penalization parameter in the objective function used to estimate the weights to better discriminate between the elements that have a weight significantly different from zero and those who have a null weight. This helps to reduce the noise produced by the elements that are not associated with the outcome which can increase the correlation between the two indices. To define the shrinkage parameter a cross-validation step should be performed. However, since this can be computationally intense, a rule of thumb that can be applied to choose the value of the parameter lambda is to set it equal to the magnitude of the AIC

of the non-penalized WQS regression. We then suggest to fit a non-penalized WQS regression with double index and then run the same model setting three different shrinkage parameter values: one equal to the magnitude of the AIC of the previous regression, one to a lower and one to a greater order of magnitude. The parameter associated to the lowest AIC will be selected. As a second step we weighted the final weight estimates by the tolerance, giving higher "importance" to the set of weights with lower collinearity between the two indices among all of those estimated in the bootstrap step. However, we strongly suggest to do a final check of the collinearity between the two scores. In the real case study we applied this new methodology taking data from the NHANES 2015-2016 study cycle to test for the association between nutrients and obesity. The results showed a harmful effect of moisture, polyunsaturated fatty acids, saturated fatty acids, sodium, caffeine and cholesterol. While some nutrients like saturated fatty acids (González-Becerra et al. 2019, Tortosa-Caparrós et al. 2017, Ralston et al. 2017, Rogero & Calder 2018), cholesterol (Tall & Yvan-Charvet 2015, Sozen & Ozer 2017) and sodium (Kang et al. 2016, Zhou et al. 2019, Lee et al. 2018) are already known risk elements for obesity, we also found nutrients for which there is a controversial evidence of their effect on BMI like moisture or caffeine. The effect of the first nutrient may depend by its intake source (Tayie & Beck 2016, Walton et al. 2019, Kant et al. 2009) and the lack of information about its origin is a limit of our study. The second element has a protective effect on a regular intake (Bhatti et al. 2013) but in excessive doses it can affect insomnia and anxiety (Nehlig 2018, Yang et al. 2010) which are associated in turn with obesity (Amiri & Behnezhad 2019, Rajan & Menon 2017, Cai et al. 2018). For this reason, the effect that we noticed could not be entirely due to caffeine intake. Because of the unavailable information of the different types of polyunsaturated fatty acids we were not able to disentangle which component drove the harmful effect on obesity. Polyunsaturated fatty acids are known to be protective against overweight and obesity (Tortosa-Caparrós et al. 2017, Ralston et al. 2017, Rogero & Calder 2018, Saini & Keum 2018, Figueiredo et al. 2018, Albracht-Schulte et al. 2018), however there is evidence that an increased intake of omega-6 long-chain polyunsaturated fatty acids can increase the risk of obesity (Saini & Keum 2018, Figueiredo et al. 2018, Albracht-Schulte et al. 2018, Fekete et al. 2015) in particular if there is an unbalanced omega-6/omega-3 ratio, an increasingly widespread problem

in western countries (Simopoulos 2016). On the other hand, a protective effect against obesity was found for beta-carotene, vitamin B12, vitamin B6, vitamin D, folic acid, vitamin C, folate DFE and alpha-carotene. For all of these nutrients there was evidence of a beneficial effect against obesity in previous studies (Coronel et al. 2019, Bonet et al. 2016, Perveen et al. 2015, Wiebe et al. 2018, Pereira-Santos et al. 2015, Savastano et al. 2017, Walsh et al. 2017, Pourshahidi 2015, Garcia-Diaz et al. 2014, Thomas-Valdés et al. 2017).
One strength of this novel approach is the ability to include all nutrients in the analysis considering the possible confounding that can be caused by the exclusion of some elements. All previous studies showed the association of one or few elements selected at a time with obesity. Moreover, we showed that considering two indices in the same WQS regression increased the accuracy of the parameter estimates when the mixture has a bidirectional effect on the outcome of interest. In addition to already available methods like the quantile-based g-computation approach, our new methodology allows to measure the double association of the mixture quantifying the effect in both the positive and negative direction with the dependent variable.

6.3 WQS for genetic data

In chapter 5 we tested the performance of WQS in a different field of application. WQS was originally developed to face the problems related to environmental exposure and the increased need of instruments able to deal with large datasets and correlated variables. This corresponds to the same need that biologists encounter when they aim to test if genes or pattern of expression affect a particular phenotype. All the available methods that are usually applied in this context require the estimate of a score in absence of the dependent variable. WQS regression has the great advantage to be a supervised method able to build an index that attributed higher weights to those genes whose down- or up-regulation shows a higher effect on the outcome. Through this study we compared the performance of WQS and ssGSEA, the most used approach in this context, and we observed how WQS gives more accurate estimates of the effect of the genetic score on the dependent variable. Moreover, WQS regression is able to identify which genes give a higher

contribution to the association with the outcome through the estimate of the weights. This can be an important additional information that allows to select the genes that can mainly regulate the considered pathway and to conduct a subsequent intervention.
The application of WQS and ssGSEA to the case study confirmed the ability of WQS in determining the biological pathway: in our case we were able to find a significant role of the genes involved in cell-cycle in the risk of death for ovarian cancer which was not observed applying ssGSEA. The measure of the effect was still less variable when using WQS regression showing narrower standard errors.

6.4 Conclusion

The advantages of WQS regression and the extension that we showed in this work are the ease of use and interpretation of the results. Moreover, the two extensions presented in this work allow to cover some problems left unsolved by other methods like Bayesian Kernel Machine Regression (BKMR) (Bobb et al. 2015), Bayesian Semiparametric Regression (BSR) (Antonelli et al. 2019) and quantile g-computation which share some common features with WQS like variable selection and measuring environmental mixture effect: none of these methods allow to consider the effect modification due to a covariate or to measure the amount of positive and negative association when the elements show both effects. On the other hand limits of this method are the lower flexibility due to the assumption of a linear trend between each element and the dependent variable compared to more flexible environmental mixture methods like BKMR and BSR and the more computational intensive procedure in contrast to methods like quantile g-computation or ssGSEA for the genetic context.
In summary, we implemented the possibility to apply the WQS regression to the generalised case when we have a binary, multinomial or count dependent variable. We then added the ability to test for an interaction between the WQS index representing the overall mixture exposure and a continuous or a categorical covariate showing the importance of considering the interaction term both during the training and validation steps. Moreover, we introduced the option to consider the bidirectionality of the mixture effect

on the outcome in the same model improving the performance of WQS to finally apply it in the genetic field, a new application context for WQS which showed promising performances. All these extensions will be implemented in the `gWQS` package which development was part of this work (Appendix A) and it is already available on CRAN.
These studies will be the starting point for additional future extensions, implementations and applications of the WQS regression.

Acknowledgements

I would like to thank my supervisors Professor Stefano Calza (University of Brescia, Italy) and Professor Monica Ferraroni (University of Milan, Italy) for their guidance and support over the last three years. I am also grateful to Professor Roberto Lucchini (Florida International University, Florida and University of Brescia, Italy) and Professor Donatella Placidi (University of Brescia, Italy) for the opportunity they gave me financially supporting my PhD and for their valuable advice. I would also express my gratitude to Dr. Chris Gennings (Icahn School of Medicine at Mount Sinai, New York) and Dr. Paul Curtin (Icahn School of Medicine at Mount Sinai, New York) for their contribution to this work and their guidance during my staying in New York. I finally want to thank my wife Elena, my children Caterina Maria and Andrea Maria, my parents, brothers and parents-in-law for their constant support.

Bibliography

Albracht-Schulte, K., Kalupahana, N. S., Ramalingam, L., Wang, S., Rahman, S. M., Robert-McComb, J. & Moustaid-Moussa, N. (2018), 'Omega-3 fatty acids in obesity and metabolic syndrome: a mechanistic update.', *The Journal of Nutritional Biochemistry* **58**, 1–16.

Amiri, S. & Behnezhad, S. (2019), 'Obesity and anxiety symptoms: a systematic review and meta-analysis.', *Neuropsychiatry* **33**(2), 72–89.

Antonelli, J., Mazumdar, M., Bellinger, D., Christiani, D. C., Wright, R. & Coull, B. (2019), 'Estimating the health effects of environmental mixtures using bayesian semiparametric regression and sparsity inducing priors.', *arXiv: Methodology* .

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., FrÃűhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., Wadlow, R. C., Le, H., Hoersch, S., Wittner, B. S., Ramaswamy, S., Livingston, D. M., Sabatini, D. M., Meyerson, M., Thomas, R. K., Lander, E. S., Mesirov, J. P., Root, D. E., Gilliland, D. G., Jacks, T. & Hahn, W. C. (2009), 'Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1', *Nature* **462**(7269), 108–12.

Bhatti, S. K., O'Keefe, J. H. & Lavie, C. J. (2013), 'Coffee and tea: perks for health and longevity?', *current Opinion in Clinical Nutrition and Metabolic Care* **16**(6), 688–97.

Bobb, J. F., Valeri, L., Henn, B. C., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J. & Coull, B. A. (2015), 'Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures.', *Biostatistics* **16**(3), 493–508.

Bonet, M. L., Canas, J. A., Ribot, J. & Palou, A. (2016), 'Carotenoids in adipose tissue biology and obesity.', *Sub-cellular Biochemistry* **79**, 377–414.

Broyden, C. G. (1970), 'The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations', *IMA Journal of Applied Mathematics* **6**(1), 76–90.

Brunst, K. J., Guerra, M. S., Gennings, C., Hacker, M., Jara, C., Enlow, M. B., Wright, R. O., Baccarelli, A. & Wright, R. J. (2017), 'Maternal lifetime stress and prenatal psychological functioning and decreased placental mitochondrial dna copy number in the prism study', *American Journal of Epidemiology* **186**(11), 1227–1236.

Cai, G.-H., Theorell-Haglow, J., Janson, C., Svartengren, M., ElmstÃěhl, S., Lind, L. & Lindberg, E. (2018), 'Insomnia symptoms and sleep duration and their combined effects in relation to associations with obesity and central obesity.', *Sleep Medicine* **46**, 81–87.

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.

CDC (2011), *Centers for Disease Control and Prevention NIfHSCN. National health and nutrition examination survey (nhanes) respiratory health spirometry procedures manual. Atlanta, GA.*

CDC (2016*a*), *National Health and Nutrition Examination Survey (NHANES). MEC In-Person Dietary Interviewers Procedures Manual.*
URL: *https://wwwn.cdc.gov/nchs/data/nhanes/2015-2016/manuals/2016_MEC_Interviewers_Procedures.pdf*

CDC (2016*b*), *National Health and Nutrition Examination Survey (NHANES). Phone Follow-Up Dietary Interviewer Procedures Manual.*
URL: *https://wwwn.cdc.gov/nchs/data/nhanes/2015-2016/manuals/2016_Phone_Follow-Up_Dietary_Interviewer_Procedures_Manual.pdf*

Coronel, J., Pinos, I. & Amengual, J. (2019), 'β-carotene in obesity re-

search: Technical considerations and current status of the field.', *Nutrients* **11**(4), 842.

Curtin, P., Kellogg, J., Cech, N. & Gennings, C. (2019), 'A random subset implementation of weighted quantile sum (wqsrs) regression for analysis of high-dimensional mixtures', *Communications in Statistics - Simulation and Computation* **0**(0), 1–16.

Czarnota, J., Gennings, C. & Wheeler, D. C. (2015), 'Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk', *Cancer Informatics* **14**(2), 159–171.

European Commission (2011), *Toxicity and Assessment of Chemical Mixtures*, European Commission, Health and Consumer Protection DirectorateâĂŞGeneral, Scientific Committee on Consumer Safety, Scientific Committee on Health and Environmental Risks, Scientific Committee on Emerging and Newly Identified Health Risks.
`https://ec.europa.eu/health/scientific_committees/environmental_risks/docs/scher_o_155.pdf` .

Fekete, K., Gyorei, E., Lohner, S., Verduci, E., Agostoni, C. & Decsi, T. (2015), 'Long-chain polyunsaturated fatty acid status in obesity: a systematic review and meta-analysis.', *Obesity Reviews* **16**(6), 488–97.

Figueiredo, P. S., Inada, A. C., Marcelino, G., Cardozo, C. M. L., de Cássia Freitas, K., de Cássia Avellaneda Guimaraes, R., de Castro, A. P., do Nascimento, V. A. & Hiane, P. A. (2018), 'Fatty acids consumption: The role metabolic aspects involved in obesity and its associated disorders.', *Nutrients* **203**, 255–267.

Fletcher, R. (1970), 'A new approach to variable metric algorithms', *The Computer Journal* **13**(3), 317–322.

Foroutan, M., Bhuva, D. D., Lyu, R., Horan, K., Cursons, J. & Davis, M. J. (2018), 'Single sample scoring of molecular phenotypes.', *BMC Bioinformatics* **19**(404).

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G., Huttenhower, C. & Waldron, L. (2013), 'curatedovariandata: clinically annotated data for the ovarian cancer transcriptome.', *Database* **2013**.

Garcia-Diaz, D. F., Lopez-Legarrea, P., Quintero, P. & Martinez, J. A. (2014), 'Vitamin c in the treatment and/or prevention of obesity.', *Journal of Nutritional Science and Vitaminology* **60**(6), 367–79.

Gennings, C., Carrico, C., Factor-Litvak, P., Krigbaum, N., Cirillo, P. M. & Cohn, B. A. (2013), 'A cohort study evaluation of maternal pcb exposure related to time to pregnancy in daughters', *Environmental Health* **12**(1).

Goldfarb, D. (1970), 'A family of variable-metric methods derived by variational means', *Mathematics of Computation* **24**(109), 23–26.

González-Becerra, K., Ramos-Lopez, Barrón-Cabrera, E., Riezu-Boj, J. I., Milagro, F. I., Martínez-López, E. & Martínez, J. A. (2019), 'Fatty acids, epigenetic mechanisms and chronic diseases: a systematic review.', *Lipids in Health and Disease* **18**(1).

Hanzelmann, S., Castelo, R. & Guinney, J. (2013), 'Gsva: gene set variation analysis for microarray and rna-seq data.', *BMC Bioinformatics* **14**(7).

Hoerl, A. E. & Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Horton, M. K., Blount, B. C., Valentin-Blasini, L., Wapner, R., Whyatt, R., Gennings, C. & Factor-Litvak, P. (2015), 'Co-occurring exposure to perchlorate, nitrate and thiocyanate alters thyroid function in healthy pregnant women', *Journal of Agricultural Biological and Environmental Statistics* **143**, 1–9.

Kang, Y. J., Wang, H. W., Cheon, S. Y., Lee, H. J., Hwang, K. M. & Yoon, H. S. (2016), 'Associations of obesity and dyslipidemia with intake of sodium, fat, and sugar among koreans: a qualitative systematic review.', *Clinical Nutrition Research* **5**(4), 290–304.

Kant, A. K., Graubard, B. I. & Atchison, E. A. (2009), 'Intakes of plain water, moisture in foods and beverages, and total water in the adult us population-nutritional, meal pattern, and body weight correlates: National health and nutrition examination surveys 1999-2006.', *The American Journal of Clinical Nutrition* **90**(3), 655–663.

Keil, A. P., Buckley, J. P., O'Brien, K. M., Ferguson, K. K., Zhao, S. & White, A. J. (2020), 'A quantile-based g-computation approach to addressing the effects of exposure mixtures.', *Environmental Health Perspective* **128**(4).

Knottnerus, J. A. & Tugwell, P. (2019), 'Confounding obscures our view, effect modification is part of reality', *Journal of Clinical Epidemiology* **114**, V – VI.

Kortenkamp, A. & Faust, M. (2018), 'Regulate to reduce chemical mixture risk', *Science* **361**(6399), 224–226.

Landrigan, P. J., Sly, J. L., Ruchirawat, M., Silva, E. R., Huo, X., Diaz-Barriga, F., Zar, H. J., King, M., Ha, E.-H., Asante, K. A., Ahanchian, H. & Sly, P. D. (2016), 'Health consequences of environmental exposures: Changing global patterns of exposure and disease', *Annals of Global Health* **82**(1), 10–9.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T. & Lee, D. (2008), 'Inferring pathway activity toward precise disease classification.', *PLOS Computational Biology* **4**(11).

Lee, J., Hwang, Y., Kim, K.-N., Ahn, C., Sung, H. K., Ko, K.-P., Oh, K.-H., Ahn, C., Park, Y. J., Kim, S., Lim, Y.-K. & Park, S. K. (2018), 'Associations of urinary sodium levels with overweight and central obesity in a population with a sodium intake.', *BMC Nutrition* **4**.

Lee, M., Rahbar, M. H., Samms-Vaughan, M., Bressler, J., Bach, M. A., Hessabi, M., Grove, M. L., Shakespeare-Pellington, S., Desai, C. C., Reece, J.-A., Loveland, K. A. & Boerwinkle, E. (2019), 'A generalized weighted quantile sum approach for analyzing correlated data in the presence of interactions.', *Biometrical Journal* **61**(4), 934–954.

Madrigal, J. M., Persky, V., Pappalardo, A. & Argos, M. (2018), 'Statistical approaches to address multi-pollutant mixtures and multiple exposures: the state of the science', *Environment International* **121**(Pt 1), 871–878.

Martin, O. V., Martin, S. & Kortenkamp, A. (2013), 'Dispelling urban myths about default uncertainty factors in chemical risk assessment - sufficient protection against mixture effects?', *Environmental Healthvolume* **12**(53).

Matthews, L. (2011), 'Cell cycle'. Provided by Reactome. Citation Accessed on Tue Oct 20 2020.
URL: *https://reactome.org/content/detail/R-HSA-1640170*

Miller, M. R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., Crapo, R., Enright, P., van der Grinten, C. P. M., Gustafsson, P.,

Jensen, R., Johnson, D. C., MacIntyre, N., McKay, R., Navajas, D., Pedersen, O. F., Pellegrino, R., Viegi, G. & Wanger, J. (2005), 'Standardisation of spirometry', *European Respiratory Society* **26**(2), 319–338.

NCHS (2017), *National Center for Health Statistics. National health and nutrition examination survey. Questionnaires, datasets, and related documentation.*
URL: *https://wwwn.cdc.gov/nchs/nhanes/Default.aspx*

Nehlig, A. (2018), 'Interindividual differences in caffeine metabolism and factors driving caffeine consumption.', *Pharmacological Reviews* **70**(2), 384–411.

Pereira-Santos, M., Costa, P. R. F., Assis, A. M. O., Santos, C. A. S. T. & Santos, D. B. (2015), 'Obesity and vitamin d deficiency: a systematic review and meta-analysis.', *Obesity Reviews* **16**(4), 341–9.

Perveen, R., Suleria, H. A. R., Anjum, F. M., Butt, M. S., Pasha, I. & Ahmad, S. (2015), 'Tomato (solanum lycopersicum) carotenoids and lycopenes chemistry; metabolism, absorption, nutrition, and allied health claims–a comprehensive review.', *Critical Reviews in Food Science and Nutrition* **55**(7), 919–29.

Pourshahidi, L. K. (2015), 'Vitamin d and obesity: current perspectives and future directions.', *The Proceeding of the Nutrition Society* **74**(2), 115–24.

Rajan, T. M. & Menon, V. (2017), 'Psychiatric disorders and obesity: A review of association studies.', *Journal of Postgraduate Medicine* **63**(3), 182–190.

Ralston, J. C., Lyons, C. L., Kennedy, E. B., Kirwan, A. M. & Roche, H. M. (2017), 'Fatty acids and nlrp3 inflammasome-mediated inflammation in metabolic tissues.', *Annual Review of Nutrition* **37**, 77–102.

Rogero, M. M. & Calder, P. C. (2018), 'Obesity, inflammation, toll-like receptor 4 and fatty acids.', *Nutrients* **10**(4).

Saini, R. K. & Keum, Y.-S. (2018), 'Omega-3 and omega-6 polyunsaturated fatty acids: Dietary sources, metabolism, and significance - a review.', *Life Sciences* **203**, 255–267.

Savastano, S., Barrea, L., Savanelli, M. C., Nappi, F., Somma, C. D., Orio, F. & Colao, A. (2017), 'Low vitamin d status and obesity: Role of nutritionist.', *Reviews in Endocrine and Metabolic Disorders* **18**(2), 215–225.

Shanno, D. F. (1970), 'Conditioning of quasi-newton methods for function minimization', *Mathematics of Computation* **24**(111), 647–656.

Simopoulos, A. P. (2016), 'An increase in the omega-6/omega-3 fatty acid ratio increases the risk for obesity.', *Nutrients* **8**(3), 128.

Sly, P. D., Carpenter, D. O., den Berg, M. V., Stein, R. T., Landrigan, P. J., Brune-Drisse, M.-N. & Suk, W. (2016), 'Health consequences of environmental exposures: Causal thinking in global environmental epidemiology', *Annals of Global Health* **82**(1), 3–9.

Sozen, E. & Ozer, N. K. (2017), 'Impact of high cholesterol and endoplasmic reticulum stress on metabolic diseases: An updated mini-review.', *Redox Biology* **12**, 456–461.

Stafoggia, M., Breitner, S., Hampel, R. & BasagaÃśa, X. (2017), 'Statistical approaches to address multi-pollutant mixtures and multiple exposures: the state of the science', *Current Environmental Health Reports* **4**(4), 481–490.

Suk, W., Ruchirawat, M., Stein, R. T., Diaz-Barriga, F., Carpenter, D. O., Neira, M. & Sly, P. D. (2016), 'Health consequences of environmental exposures in early life: Coping with a changing world in the post-mdg era', *Annals of Global Health* **82**(1), 20–7.

Tall, A. R. & Yvan-Charvet, L. (2015), 'Cholesterol, inflammation and innate immunity.', *Nature Reviews Immunology* **15**(2), 104–16.

Tanner, E. M., Bornehag, C.-G. & Gennings, C. (2019), 'Repeated holdout validation for weighted quantile sum regression', *MethodsX* **6**, 2855 – 2860.

Tayie, F. A. & Beck, G. L. (2016), 'Alcoholic beverage consumption contributes to caloric and moisture intakes and body weight status.', *Nutrition* **32**(7-8), 799–805.

Thomas-Valdés, S., das Gracas V Tostes, M., Anunciacao, P. C., da Silva, B. P. & Sant'Ana, H. M. P. (2017), 'Association between vitamin deficiency and metabolic disorders related to obesity.', *Critical Reviews in Food Science and Nutrition* **57**(15), 3332–3343.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Tomfohr, J., Lu, J. & Kepler, T. B. (2005), 'Pathway level analysis of gene expression using singular value decomposition.', *BMC Bioinformatics* **6**(225).

Tortosa-Caparrós, E., Navas-Carrillo, D., Marín, F. & Orenes-Pinero, E. (2017), 'Anti-inflammatory effects of omega 3 and omega 6 polyunsaturated fatty acids in cardiovascular disease and metabolic syndrome.', *Critical Review in food Science and Nutrition* **57**(16), 3421–3429.

Tu, Y.-K., Gunnell, D. & Gilthorpe, M. S. (2008), 'Simpson's paradox, lord's paradox, and suppression effects are the same phenomenon - the reversal paradox', *Emerging Themes in Epidemiology* **5**(2).

Walsh, J. S., Bowles, S. & Evans, A. L. (2017), 'Vitamin d in obesity.', *Current Opinion in Endocrinology, Diabetes and Obesity* **24**(6), 389–394.

Walton, J., O'Connor, L. & Flynn, A. (2019), 'Cross-sectional association of dietary water intakes and sources, and adiposity: National adult nutrition survey, the republic of ireland.', *European Journal of Nutrition* **58**(3), 1193–1201.

Wiebe, N., Field, C. J. & Tonelli, M. (2018), 'A systematic review of the vitamin b12, folate and homocysteine triad across body mass index.', *Obesity Reviews* **19**(11), 919–29.

Yang, A., Palmer, A. A. & de Wit, H. (2010), 'Genetics of caffeine consumption and responses to caffeine.', *Psychopharmachology* **211**(3), 245–57.

Zhou, L., Stamler, J., Chan, Q., Horn, L. V., Daviglus, M. L., Dyer, A. R., Miura, K., Okuda, N., Wu, Y., Ueshima, H., Elliott, P. & Zhao, L. (2019), 'Salt intake and prevalence of overweight/obesity in japan, china, the united kingdom, and the united states: the intermap study.', *American Journal of Clinical Nutrition* **110**(1), 34–40.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the Royal Statistical Society. Series B (Methodological)* **101**(476), 1418–1429.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society. Series B (Methodological)* **67**(2), 301–320.

Appendix A

The gWQS *R* package

The *R* package gWQS extends WQS, WQS_{RS} and WQS_{RH} regression to applications with continuous, categorical and count outcomes. The first main function of the gWQS package is gwqs, which allows the implementation of WQS and WQS_{RS} regression, while the function gwqsrh that relies on the gwqs function, allows to apply the WQS_{RH} method. For all the three methodologies a linear, logistic, multinomial, Poisson, quasi-Poisson and negative binomial regression are implemented. For Poisson and negative binomial regression a zero inflated option is also implemented. Few secondary functions are also available to generate plots and tables from the gwqs and gwqsrh output and will be described in the following four examples. The gWQS package uses the optim function from the stats package as optimization algorithm to estimates the weights. This function allows to solve general nonlinear programming problems through the Broyden, Fletcher, Goldfarb and Shanno (BFGS) method (Broyden 1970, Fletcher 1970, Goldfarb 1970, Shanno 1970), a quasi-Newton method also known as a variable metric algorithm. A quadratic transformation is applied to the weights while optimizing the objective function in order to constrain them to be positive; they are then normalized to get their sum equal to one.
We created the wqs_data dataset (available once the package is installed and loaded) to demonstrate the use of these functions. These data reflect 59 exposure concentrations simulated from a distribution of 34 PCB exposures and 25 phthalate biomarkers measured in subjects participating in the NHANES study (2001-2002). Additionally, 8 outcome measures were simulated applying different distributions and fixed beta coefficients to the predictors. In particular y and yLBX were simulated from a normal distribution, ybin and ybinLBX from a binomial distribution, ymultinom and ymultinomLBX from a

multinomial distribution and `ycount` and `ycountLBX` from a Poisson distribution. Table A.1 shows the real beta coefficient values used to generate the dependent variables:

	yLBX, ybinLBX, ycountLBX	y, ybin, ycount	ymultinomLBX		ymultinom	
			Level B	Level C	Level B	Level C
PCBs						
LBX138LA	0.6	0.6	0.8	0	0.8	0
LBXD02LA	0.45	0.45	0	0.6	0	0.6
LBXF07LA	0.45	0.45	0	0.6	0	0.6
LBX105LA	0.3	0.3	0	0.4	0	0.4
LBXF06LA	0.3	0.3	0	0.4	0	0.4
LBX157LA	0.2	0.2	0	0.3	0	0.3
LBXD04LA	0.15	0.15	0.2	0	0.2	0
Phthalates						
URXMOH	0	0.45	0	0	0	0.6
URXP10	0	0.3	0	0	0.4	0
URXP02	0	0.2	0	0	0.3	0
URXUCR	0	0.2	0	0	0.3	0
URXMC1	0	0.15	0	0	0.2	0

Table A.1: Real beta coefficient values used to generate the dependent variables.

The `sex` variable was also simulated to allow to adjust for a covariate in the model. This dataset can thus be used to test the `gWQS` package by analyzing the mixture effect of the 59 simulated chemicals on the outcomes, with adjustments for covariates.
We list five examples to illustrate the usage of the package.

A.1 Example 1

The following script calls a WQS model for a continuous outcome using the function `gwqs` that returns an object of class `gwqs`; the three functions `gwqs_barplot`, `gwqs_scatterplot` and `gwqs_fitted_vs_resid` allow to plot the figures shown in figure A.1, while the functions `gwqs_summary_tab` and `gwqs_weights_tab` allow to generate the summary and weights tables:

```
R> library(gWQS)
R> # we save the names of the mixture variables in the variable "PCBs"
R> PCBs <- names(wqs_data)[1:34]
R> # we run the model and save the results in the variable "results"
R> results <- gwqs(yLBX ~ wqs, mix_name = PCBs,
+                  data = wqs_data, q = 10,
```

```
+                     validation = 0.6, b = 100,
+                     b1_pos = TRUE, b1_constr = FALSE,
+                     family = "gaussian", seed = 2016)
R> # bar plot
R> gwqs_barplot(results)
R> # scatter plot y vs wqs
R> gwqs_scatterplot(results)
R> # scatter plot residuals vs fitted values
R> gwqs_fitted_vs_resid(results)
```

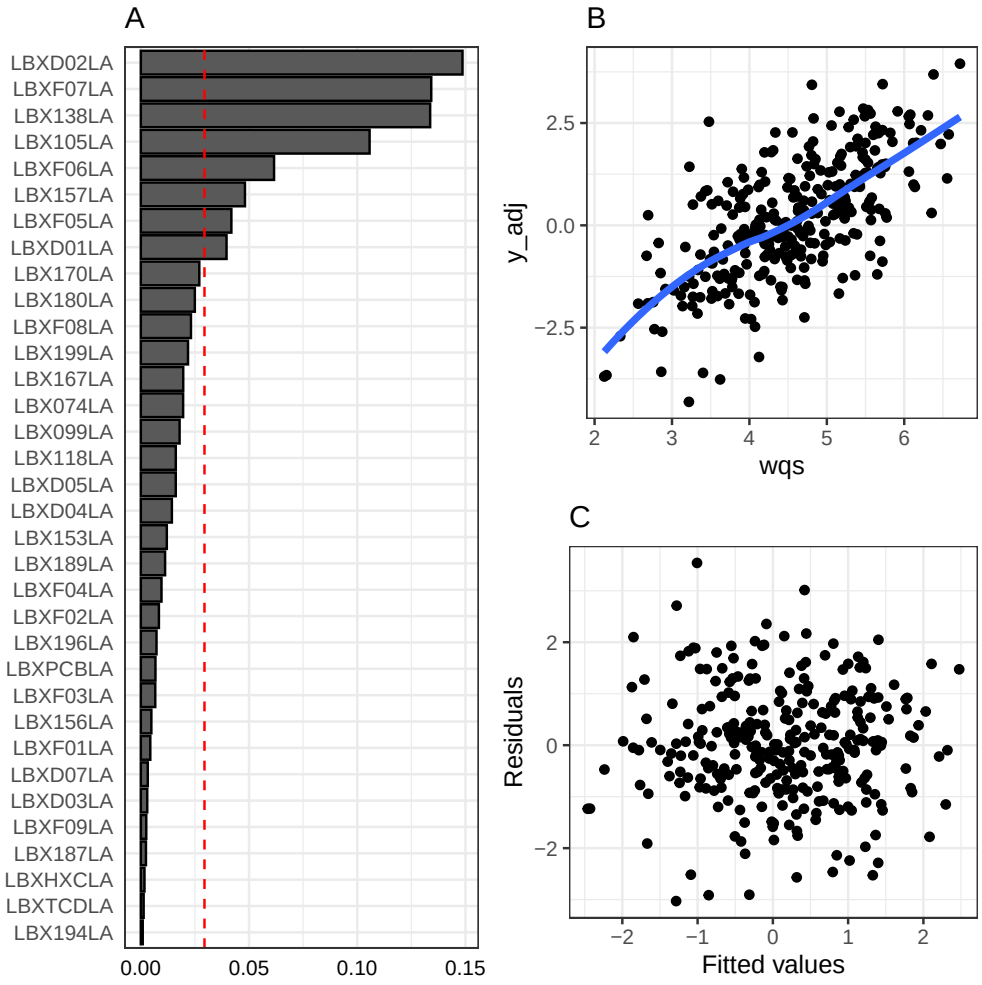


Figure A.1: Plots available for linear outcomes

This WQS model tests the relationship between our dependent variable, `yLBX`, and a WQS index estimated from ranking exposure concentrations in deciles (`q = 10`); in the `gwqs` formula the `wqs` term must be included as if a `wqs` variable was present in the dataset. The data were divided in 40% of the dataset for training and 60% for validation (`validation = 0.6`), and

100 bootstrap samples (`b = 100`) for parameter estimation were assigned (in practical applications we suggest at least 100 bootstrap samples to be used). Because WQS provides a unidirectional evaluation of mixture effects, we first examined weights derived from bootstrap models where β_1 was positive (`b1_pos = TRUE`); we could test for negative associations by setting that parameter to be false (`b1_pos = FALSE`). We can also choose to constrain the β_1 to be positive (`b1_pos = TRUE` and `b1_constr = TRUE`) or negative (`b1_pos = FALSE` and `b1_constr = TRUE`) when we estimate the weights; in the case of example 1 we are not applying a constraint to β_1. We linked our model to a gaussian distribution to test for relationships between the continuous outcome and exposures (`family = "gaussian"`), and fixed the seed to 2016 for reproducible results (`seed = 2016`).
To test the statistical significance of the association between the variables in the model, the following code has to be run as for a classical R regression function:

```
R> summary(results$fit)

Call:
glm(formula = formula, family = family, data = bdtf, weights = wghts)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0285  -0.6612  -0.0078   0.6998   3.5401

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.74268    0.32599  -14.55   <2e-16 ***
wqs          1.07415    0.07023   15.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.196741)

    Null deviance: 636.62  on 299  degrees of freedom
Residual deviance: 356.63  on 298  degrees of freedom
AIC: 909.24

Number of Fisher Scoring iterations: 2
```

This result tells us that the association between the WQS index and the outcome $yLBX$ is positive and statistically significant (`p<2e-16`).
Figure A.1 A is a barplot showing the weights assigned to each variable or-

dered from the highest weight to the lowest. The dashed red line represents the cutoff τ to discriminate which element has a significant weight greater than zero. As suggested in Carrico et al. (2015) the default value is set to $\tau = 1/c$ where c is the number of elements in the mixture (in our case $c = 34$). From these results we notice that the variables `LBXD02LA, LBXF07LA, LBX138LA, LBX105LA, LBXF06LA, LBX157LA, LBXF05LA` and `LBXD01LA` are the largest contributors to the mixture effect. According to table A.1 the model was able to find all the elements associated with the outcome apart from `LBXD04LA` and wrongly attributed a high weight to `LBXF05LA` and `LBXD01LA`.

To have the exact values of the estimated weights we can apply the command `results$final_weights`. The following code shows the first six highest weights; the full list of weights can be called by omitting the head function:

```
R> head(results$final_weights)
```

```
         mix_name mean_weight
LBXD02LA LBXD02LA  0.14862139
LBXF07LA LBXF07LA  0.13418317
LBX138LA LBX138LA  0.13368244
LBX105LA LBX105LA  0.10570102
LBXF06LA LBXF06LA  0.06156063
LBX157LA LBX157LA  0.04811495
```

This same table as well as the summary table are also shown in the Viewer window through the functions `gwqs_weights_tab` and `gwqs_summary_tab` respectively. Both these two functions use the package `kableExtra` to produce the output. The output (table A.2 and A.3) and respective code is shown below (only the first 10 elements with highest weights are reported in table A.3):

```
R> # summary table
R> gwqs_summary_tab(results)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.74	0.3260	-14.5	0
wqs	1.07	0.0702	15.3	0

Table A.2: Summary results of the WQS regression for linear outcomes.

```
R> mf_df = as.data.frame(signif(coef(summary(results$fit)), 3))
R> kable_styling(kable(mf_df, row.names = TRUE))
```

mix_name	mean_weight
LBXD02LA	0.149000
LBXF07LA	0.134000
LBX138LA	0.134000
LBX105LA	0.106000
LBXF06LA	0.061600
LBX157LA	0.048100
LBXF05LA	0.041800
LBXD01LA	0.039600
LBX170LA	0.027000
LBX180LA	0.025000

Table A.3: Weights table of the WQS regression for linear outcomes.

```
R> # weights table
R> gwqs_weights_tab(results)
```

```
R> final_weight <- results$final_weights
R> final_weight[, -1] <- signif(final_weight[, -1], 3)
R> kable_styling(kable(final_weight, row.names = FALSE))
```

In plot B of figure A.1 we have a representation of the wqs index and the outcome (adjusted for the model residual when covariates are included in the model) that shows the direction and the shape of the association between the exposure and the outcome. For example, in this case we can observe a linear and positive relationship between the mixture and the `yLBX` variable.
In plot C a diagnostic graph of the residuals vs the fitted values is shown to check if they are randomly spread around zero or if there is a trend. All these plots are built using the `ggplot2` package.
The `gwqs` function gives back other outputs like the vector of the values that indicate whether the solver has converged (0) or not (1) (`results$conv`), the matrix with all the estimated weights and the associated β_1, standard errors, statistics and p-values for each bootstrap sample (`results$bres`), the vector of the estimated `wqs` index (`results$wqs`), the list of vectors containing the cutoffs used to determine the quantiles of each variable in the mixture (`results$qi`), the list of vectors containing the rows of the subjects included in each bootstrap dataset (`results$bindex`), the rows identifying the subjects used to estimate the weights in each bootstrap (`results$tindex`), the rows identifying the subjects used to estimate the parameters of the final model (`results$vindex`), the vector of the values of the objective function at the optima parameter estimates obtained at each bootstrap step (`results$objfn_values`) and any messages from the `optim` function (`resul-`

`ts$optim_messages`).
The following script allows to reproduce the figures that are automatically generated using the plots functions:

```
R> # bar plot
R> w_ord <- order(results$final_weights$mean_weight)
R> mean_weight <- results$final_weights$mean_weight[w_ord]
R> mix_name <- factor(results$final_weights$mix_name[w_ord],
+                     levels = results$final_weights$mix_name[w_ord])
R> data_plot <- data.frame(mean_weight, mix_name)
R> nPCBs <- length(PCBs)
R> ggplot(data_plot, aes(x = mix_name, y = mean_weight)) +
+    geom_bar(stat = "identity", color = "black") + theme_bw() +
+    theme(axis.ticks = element_blank(),
+          axis.title = element_blank(),
+          axis.text.x = element_text(color='black'),
+          legend.position = "none") + coord_flip() +
+    geom_hline(yintercept = 1/nPCBs, linetype="dashed", color = "red")
R> #
R> # scatter plot y vs wqs
R> ggplot(results$y_wqs_df, aes(wqs, y_adj)) + geom_point() +
+    stat_smooth(method = "loess", se = FALSE, size = 1.5) + theme_bw()
R> #
R> # scatter plot residuals vs fitted values
R> fit_df <- data.frame(fitted = fitted(results),
+                       resid = residuals(results, type = "response"))
R> ggplot(fit_df, aes(x = fitted, y = resid)) + geom_point() +
+    theme_bw() + xlab("Fitted values") + ylab("Residuals")
```

A.2 Example 2

In the following code we run a logistic regression (`family = binomial`) to test the association between the exposure to all the 59 elements in the mixture and the outcome `ybin`. In this case we apply the WQS_{RS} method (`rs = TRUE`) creating 1000 random subsets (`b = 1000`) selecting without replacement $\sqrt{(c)}$ (as default) variables from the entire set (in our case each random subset contains 8 variables, a different number can be chosen through the parameter `n_vars`). An exponential signal function is applied to the t-statistic associated to the WQS parameter and then used as weight when averaging the final WQS weights in order to give more importance to the estimates of those weights related to a more significant β parameter. A quadratic (default) or an exponential signal function is recommended especially when using the

WQS_{RS}. Since the mixture concentrations in this example are already standardized we can also run a model without categorizing for quantiles (`q = NULL`) after checking that there were no skewed distributions. Furthermore we examined the ability of our model to predict the outcome on a third part of the dataset: we split the data in a training dataset (`wqs_data_train`) used to build the model and a second part was used for prediction (`wqs_data_pred`). The training dataset was further split in test and validation (40% and 60% of the training dataset respectively (`validation = 0.6`)) to build the WQS_{RS} model.

```
R> # we save the names of the mixture variables in the variable
R> # "toxic_chems"
R> toxic_chems <- names(wqs_data)[1:59]
R> # create a dataset exluding the data where we want to apply the
R> # prediction and define the group variable to identify the test
R> # and validation dataset
R> set.seed(1234)
R> pred <- sample(1:nrow(wqs_data), 150)
R> wqs_data_train <- wqs_data[-pred,]
R> wqs_data_pred <- wqs_data[pred,]
R> # we run the logistic model and save the results in the variable
R> # "results2"
R> results2 <- gwqs(ybin ~ wqs, mix_name = toxic_chems, rs = TRUE,
+                   signal = "exp", data = wqs_data_train, q = 10,
+                   validation = 0.6, b = 1000, b1_pos = TRUE,
+                   b1_constr = FALSE, family = binomial, seed = 2016)
R> # bar plot
R> gwqs_barplot(results2)
R> # scatter plot ybin vs wqs
R> gwqs_scatterplot(results2)
R> # plot ROC
R> gwqs_ROC(results2, wqs_data_pred)
```

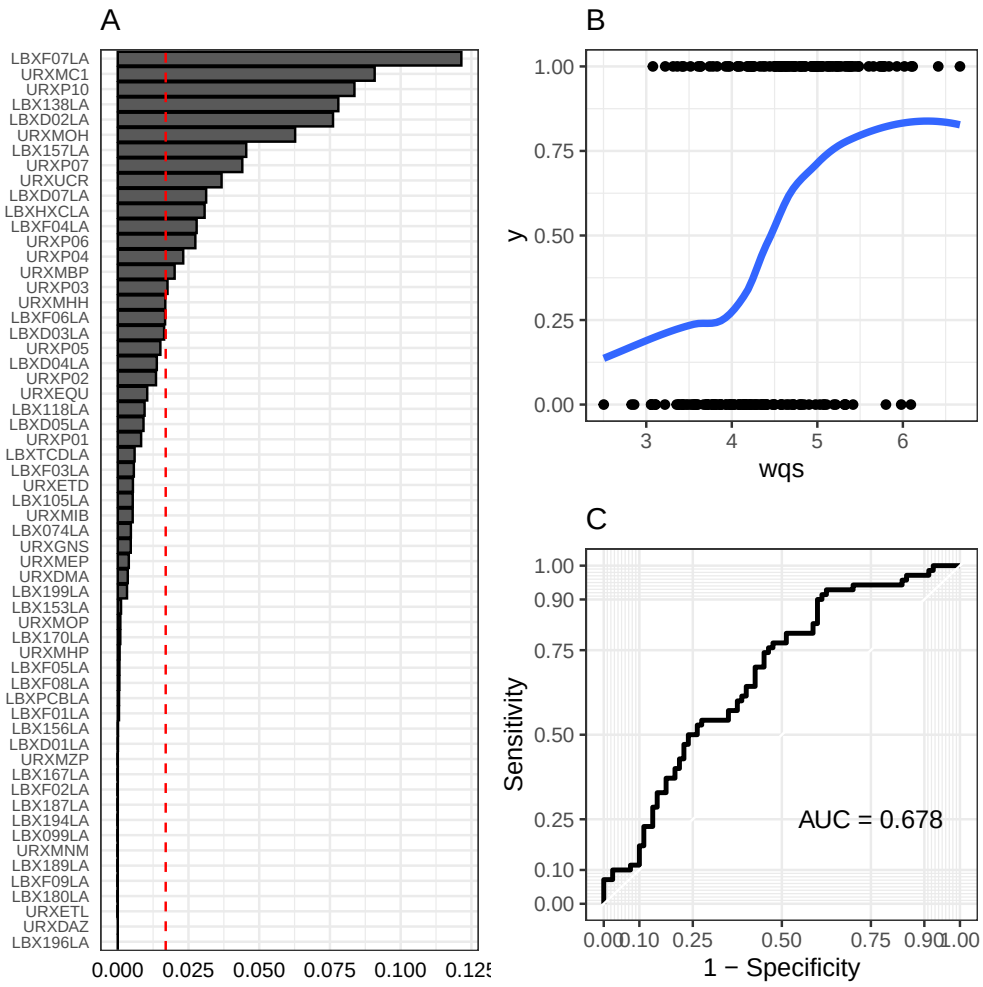


Figure A.2: Plots available for binary outcomes

As we can see from table A.4 (generated by the function `gwqs_summary_tab`) there is a statistically significant association between the WQS index and the outcome `ybin` (p=1.47e-07):

```
R> # summary table
R> gwqs_summary_tab(results2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.00546	0.152	-0.036	0.971
wqs	3.37000	0.608	5.550	0.000

Table A.4: Summary results of the WQS_{RS} regression for binary outcomes.

Figure A.2 A and B are generated by the same functions `gwqs_barplot`, `gwqs_scatterplot` as reported in the code above. From figure A.2 A we see the per-variable calculated weights, ordered by relative magnitude. All

the elements associated with the dependent variable were correctly identified apart from LBX105LA, while there were few non significant weights wrongly included: LBXD07LA, URXP07, URXP06, URXP04, LBXHXCLA and URXP01. Plot B shows a positive relationship between the mixture and the outcome confirming the results in table A.4. Through the predict function we can run the predictive model. The following code shows how to reproduce the prediction: the function predict requires the object results2 of class gwqs and the optional new dataset (argument newdata) on which the prediction model is applied; alternatively if the newdata argument is not specified the predictive model is applied to the same data on which the fitted model was built. The predict function returns the dataset df_pred, which is a data.frame including a first column as the actual value of the dependent variable and a second column as the predicted values, the matrix Q of the elements in the mixture categorized in their quantiles, the list qi containing the vectors of the cut points used to generate the quantile variables and the vector wqs representing the WQS index built using the new data. The following code shows how to apply the predict function on the gwqs output:

```
R> results2_pred <- predict(results2, newdata = wqs_data_pred)
```

For the predictive logistic regression model the gWQS provides the function gwqs_ROC to plot the Receiver Operating Characteristic (ROC) curve. Figure A.2 C shows the ROC curve related to the predictive model: we can see that the cutoff that is closer to the left-hand border and the top border has around 70% sensitivity (true positive) and 70% $1-specificity$ (false positive) while the Area Under the Curve (AUC) is equal to 0.687.
The same plot as in figure A.2 C can be displayed through the code after having installed and loaded the package plotROC:

```
R> # plot ROC curve
R> gg_roc <- ggplot(results2_pred$df_pred, aes(d=y, m=ypred)) +
+    geom_roc(n.cuts = 0) +
+    style_roc(xlab = "1 - Specificity", ylab = "Sensitivity")
R> auc_est <- plotROC::calc_auc(gg_roc)
R> gg_roc + annotate("text", x=0.75, y=0.25,
+                    label=paste0("AUC = ", round(auc_est[, "AUC"], 3)))
```

A.3 Example 3

In this third case we fit a WQS multinomial model (`family = "multinomial"`) for categorical data: the outcome is `ymultinomLBX` consisting of three categories `"A", "B"` and `"C"`. This modelling strategy creates a distinct logistic model comparing each level of the outcome variable to a reference level (in this case the `"A"` category). We chose to create the training and validation dataset and assign to `valid_var` the name of the variable that identifies the two datasets (`valid_var = "group"`). In this case we had to choose two directions for each level of the outcome variable (in this case both positive: `b1_pos = c(TRUE, TRUE)`). We also decided to run the bootstrap in parallel on multiple cores (`plan_strategy = "multisession"`) through the `future_lapply` function from the package `future.apply`.

```
R> # we create the variable "group" in the dataset to identify the
R> # training and validation dataset: we choose 300 observations for
R> # the validation dataset and the remaining 200 for the training dataset
R> set.seed(123)
R> wqs_data$group <- 0
R> wqs_data$group[rownames(wqs_data) %in%
+                   sample(rownames(wqs_data), 300)] <- 1
R> #
R> # we run the logistic model and save the results in the variable
R> # "results3"
R> results3 <- gwqs(ymultinomLBX ~ wqs, mix_name = PCBs,
+                   data = wqs_data, q = NULL, valid_var = "group",
+                   b = 100, b1_pos = c(TRUE, TRUE), b1_constr = FALSE,
+                   family = "multinomial", seed = 123,
+                   plan_strategy = "multisession")
R> # bar plot
R> gwqs_barplot(results3)
R> # scatter plot y vs wqs
R> gwqs_scatterplot(results3)
R> # weights scatterplot
R> gwqs_levels_scatterplot(results3)
```

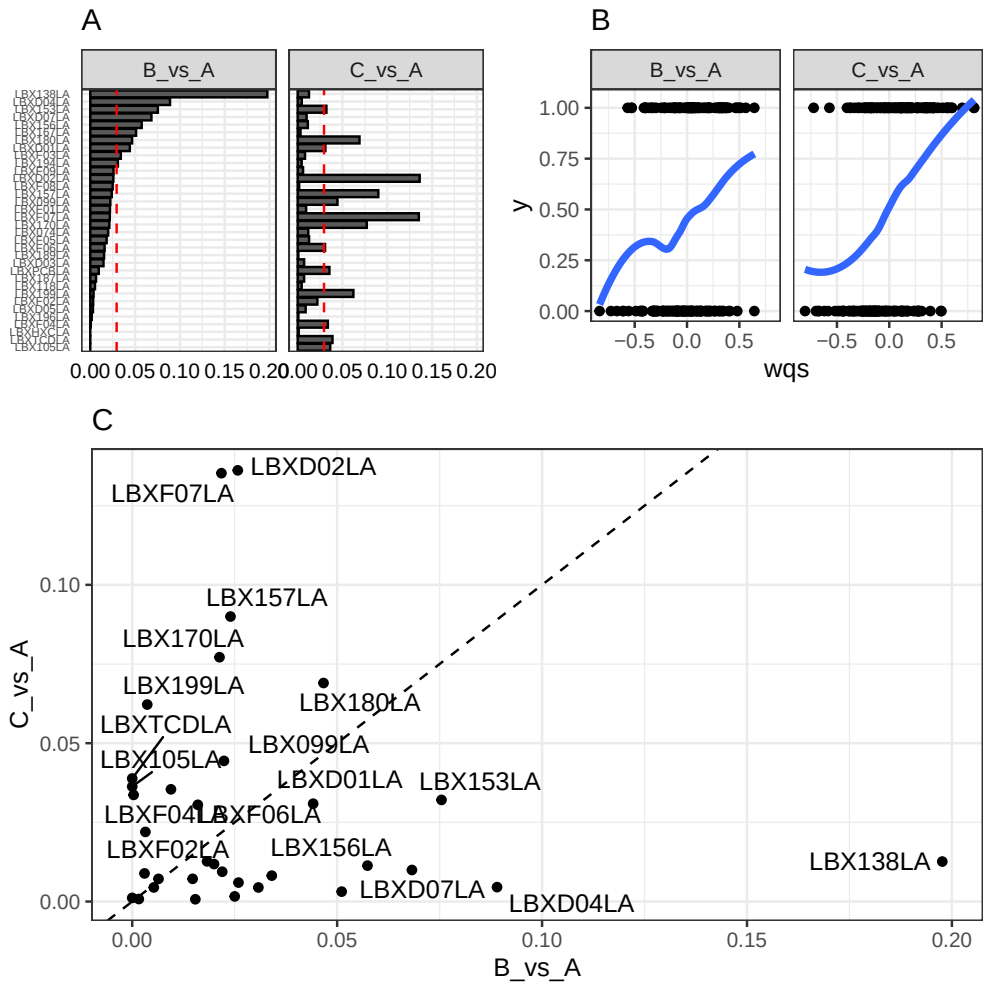


Figure A.3: Plots available for multinomial outcomes

To test for significance we still can apply the `gwqs_summary_tab` function:

```
R> # summary table
R> gwqs_summary_tab(results3)
```

	Estimate	Standard_Error	stat	p_value
(Intercept)_B_vs_A	-0.1380	0.149	-0.93	0.352000
wqs_B_vs_A	1.7800	0.495	3.60	0.000324
(Intercept)_C_vs_A	0.0486	0.147	0.33	0.742000
wqs_C_vs_A	3.1300	0.501	6.24	0.000000

Table A.5: Summary results of the WQS regression for multinomial outcomes.

As we can see from the results in table A.5, both the WQS indices for each level are significant ($p < 0.001$) but, as shown from plot A and C in figure A.3, chemicals have different weights depending on the race. The first ten highest

weights ordered for category "B are reported in table A.6 obtained through the function gwqs_weights_tab:

```
R> # weights table
R> gwqs_weights_tab(results3)
```

mix_name	B_vs_A	C_vs_A
LBX138LA	1.91e-01	0.006720
LBXD04LA	9.32e-02	0.003090
LBX153LA	7.88e-02	0.035700
LBX180LA	7.43e-02	0.068200
LBX167LA	7.01e-02	0.003370
LBXD07LA	5.67e-02	0.008220
LBXF03LA	5.00e-02	0.011100
LBX156LA	4.47e-02	0.009110
LBXD01LA	3.63e-02	0.031400
LBXF07LA	2.95e-02	0.141000

Table A.6: Weights table of the WQS regression for multinomial outcomes.

In figure A.3 while plots A and B are the same as in figure A.1 and A.2 generated by the same functions (gwqs_barplot and gwqs_scatterplot) but divided by the levels of the outcome variable, C is a scatter plot of the weights generated by the function gwqs_levels_scatterplot. This allows us to compare the magnitude of weights estimated in each model (e.g. "B vs A" or "C vs A"), with departures from the main diagonal indicating variables that are differentially-weighted for each comparison, e.g. C vs. A, or B vs. A. In our case we can see that there is a clear discrepancy between the levels since the highest weights in one level have lower weights in the other level and the other way round. This is plotted only when the outcome has three levels.
The following code shows how to generate the plots in the multinomial case as shown by figure A.3:

```
R> # bar plot
R> data_plot <- results3$final_weights[order(results3$final_weights[,2]),]
R> pos <- match(data_plot$mix_name, sort(data_plot$mix_name))
R> data_plot$mix_name <- factor(data_plot$mix_name,
+                               levels(data_plot$mix_name)[pos])
R> data_plot_l <- melt(data_plot, id.vars = "mix_name")
R> ggplot(data_plot_l, aes(x = mix_name, y = value)) +
+    facet_wrap(~ variable) +
+    geom_bar(stat = "identity", color = "black") +
+    theme_bw() + theme(axis.ticks = element_blank(),
+                       axis.title = element_blank(),
```

```
+                       axis.text.x = element_text(color='black'),
+                       legend.position = "none") + coord_flip() +
+    geom_hline(yintercept = 1/nPCBs, linetype="dashed", color = "red")
R> #
R> # scatter plot y vs wqs
R> ggplot(results3$y_wqs_df, aes(wqs, y)) + geom_point() +
+    stat_smooth(method = "loess", se = FALSE, size = 1.5) +
+    theme_bw() + facet_wrap(~ level)
R> #
R> # scatter plot of weights for the two levels of the dependent variable
R> ggplot(data_plot, aes_string(names(data_plot)[2],
+                               names(data_plot)[3])) +
+    geom_point() + theme_bw() + xlab(names(data_plot)[2]) +
+    ylab(names(data_plot)[3]) + geom_abline(linetype = 2) +
+    ggrepel::geom_text_repel(aes(label=mix_name))
```

A.4 Example 4

This last example shows how to fit a WQS_{RH} regression, in particular we apply the method on count data. The dependent variable taken into account is `ycountLBX` and we fit a Poisson regression (`family = poisson`). A quasi-Poisson regression is also available when count data are overdispersed (`family = quasipoisson`). In this case we need to specify the number of repeated holdout validation sets and we set it to 10 (`rh = 10`) as well as the number of bootstrap (`b = 10`) to reduce the computation time. However, we suggest to run the WQS_{RH} with at least 100 repeated holdout and 100 bootstrap. We also run a stratified analysis by sex estimating different weights for males and females setting `stratified = "sex"`.

```
R> # we run the poisson model and save the results in the variable
R> # "results4"
R> results4 <- gwqsrh(ycountLBX ~ wqs + sex, mix_name = PCBs, rh = 10,
+                     stratified = "sex", data = wqs_data, q = 10,
+                     validation = 0.6, b = 10, b1_pos = TRUE,
+                     b1_constr = FALSE, family = poisson, seed = 123)
R> # box plot
R> gwqsrh_boxplot(results4)
R> # bar plot
R> gwqs_barplot(results4)
R> # scatter plot y vs wqs
R> gwqs_scatterplot(results4)
```

```
R> # scatter plot residuals vs fitted values
R> gwqs_fitted_vs_resid(results4)
```

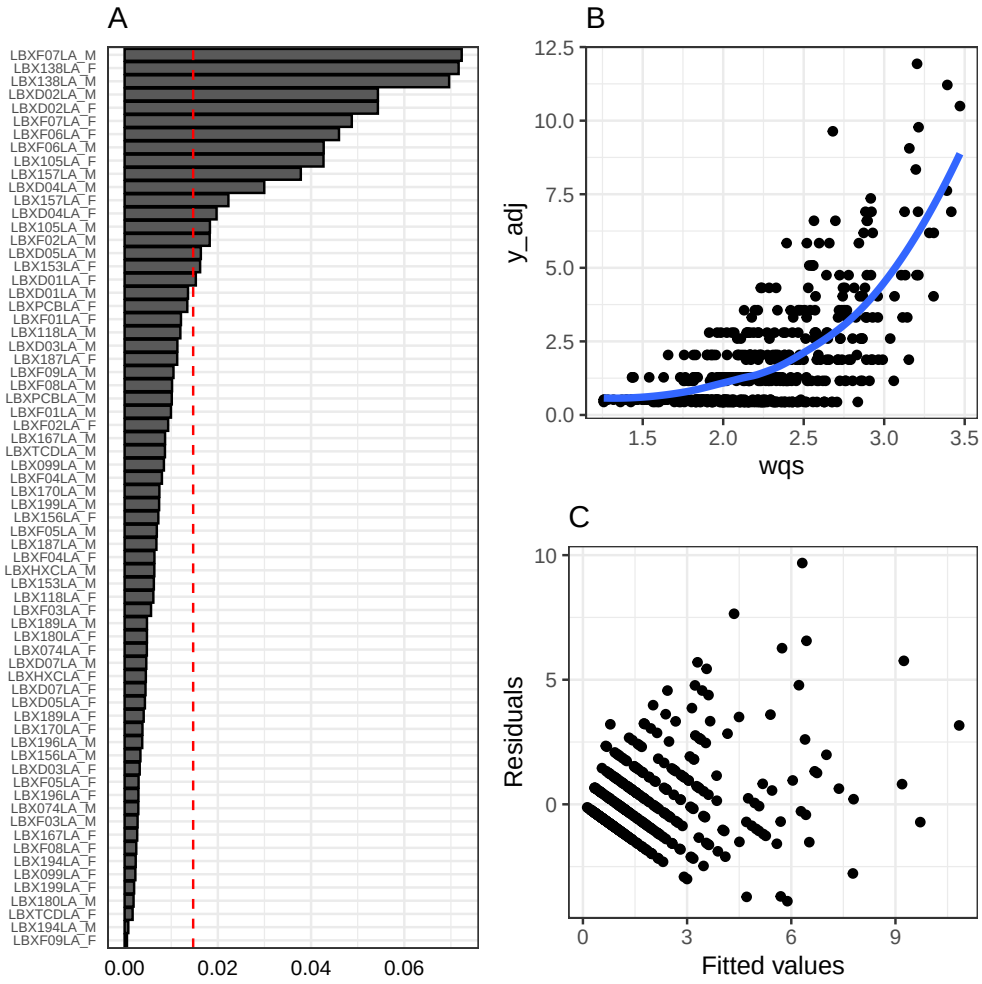


Figure A.4: Plots available for count outcome

The results of the model are shown in table A.7 and table A.8. When using the WQS_{RH} method the additional option `sumtype` is available when applying the method functions (such as `summary`, `predict`, `residuals` etc.) or the secondary functions to create plots and tables. Through this option we can choose if using the mean and the 95% CI based on the standard deviation (`sumtype = "norm"`, default value) or the median and the 2.5^{th} and 97.5^{th} percentiles (`sumtype = "perc"`). As a default the mean and standard deviation based 95% CI are set:

```
R> # summary table
R> gwqs_summary_tab(results4)
```

	Estimate	Std. Error	2.5 %	97.5 %
(Intercept)	-4.640	0.4140	-5.460	-3.83
wqs	2.020	0.0955	1.840	2.21
sexF	0.687	0.5080	-0.309	1.68

Table A.7: Summary results of the WQS_{RH} regression for Poisson regression.

```
R> # weights table
R> gwqs_weights_tab(results4)
```

mix_name	Estimate	2.5 %	97.5%
LBXF07LA_M	0.072300	5.21e-02	0.08330
LBX138LA_F	0.071700	5.93e-02	0.08240
LBX138LA_M	0.069600	5.22e-02	0.08420
LBXD02LA_F	0.054300	3.90e-02	0.06480
LBXD02LA_M	0.054300	3.12e-02	0.07460
LBXF07LA_F	0.048700	3.42e-02	0.06420
LBXF06LA_F	0.046000	3.06e-02	0.06960
LBXF06LA_M	0.042700	2.59e-02	0.06170
LBX105LA_F	0.042700	1.91e-02	0.06160
LBX157LA_M	0.037800	8.01e-03	0.05640

Table A.8: Weights table of the WQS_{RH} regression for Poisson regression.

We notice that there is a significant positive association between the WQS index and the dependent variable (table A.7 and figure A.4 B). Since we stratified by sex, we have an estimate of each weight for males and females and we can see how the weights differ between the two genders (table A.8 and figure A.4 A): we have a good agreement between the first four couple of weights (e.g. `LBXF07LA, LBX138LA, LBXD02LA` and `LBXF06LA` has an high impact in both males and females).
Since in WQS_{RH} we repeat a WQS regression `rh` times, the additional function `gwqsrh_boxplot(results4)` is available allowing to build the plot reported in figure A.5.

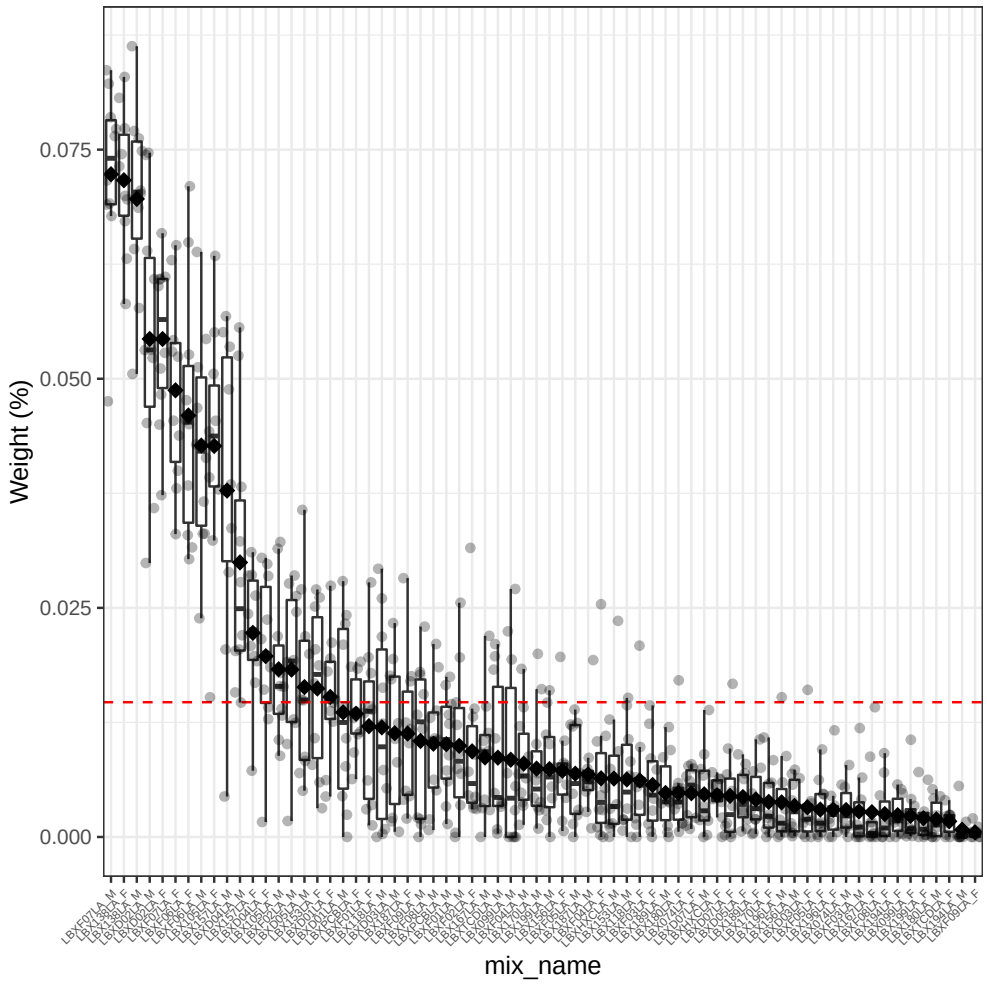


Figure A.5: Box Plot of the weights in WQS_{RH}.

The box plots represent the weight distributions for each element in the mixture and the estimated mean value (represented by the diamond). The red dashed line is the prespecified cutoff τ to determine the significant weights. An additional advantage of this method is the possibility to look at the variability of each weight and the ability to overcome the risk of incorrect conclusions that could incur in single partition analysis. Below is reported the code to reproduce the same plot as in figure A.5.

```
R> # box plot
R> wboxplot <- melt(results4$wmat, varnames = c("rh", "mix_name"))
R> wboxplot$mix_name <- factor(wboxplot$mix_name,
+                              levels = results4$final_weights$mix_name)
R> ggplot(wboxplot, aes_string(x = "mix_name", y = "value")) +
+    geom_boxplot(outlier.shape = " ") + theme_bw() +
+    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
```

```
+    ylab("Weight (%)") +
+    stat_summary(fun.y = mean, geom = "point", shape = 18, size = 3) +
+    geom_hline(yintercept=1/(2*nPCBs),linetype="dashed",color="red") +
+    geom_jitter(alpha = 0.3)
```

A zero-inflated model can be fitted for the Poisson and negative binomial regression setting `zero_infl = TRUE` and choosing a link function for the binomial process (we can choose among `"logit"`, `"probit"`, `"cloglog"`, `"cauchit"`, and `"log"`).