Workshop on Analyzing Mixtures in Environmental Health Studies: WQS Regression

Chris Gennings

Icahn School of Medicine at Mount Sinai
Department of Environmental Medicine and Public Health

August 24, 2018

Icahn
School of
Medicine at
Mount
Sinai

Overview of Mixtures

Concerns: high dimensionality; complex correlation patterns

- multicollinearity and
- reversal paradox
- Sensitivity and specificity identifying ‘bad actors’

Strategies:

- Reducing dimensionality: e.g., PCA
- Addressing ill-conditioning in regression with constraints
  - Shrinkage methods – e.g., LASSO
  - WQS regression
- Flexible response surface methods
  - e.g., Bayesian Kernel Machine Regression (BKMR)

Multicollinearity

- Correlation among predictor variables impact the variability of parameter estimates in regression models.

- The prediction of the model at observed data points may be adequate (i.e., “the old picket fence” analogy), but hypothesis tests of model parameters have decreased power.

Reversal paradox

Illustration: Assume Corr(y, x1)=0.2 and Corr(y, x2)=0.1. The beta estimates in a linear model are impacted by the Corr(X1, X2):

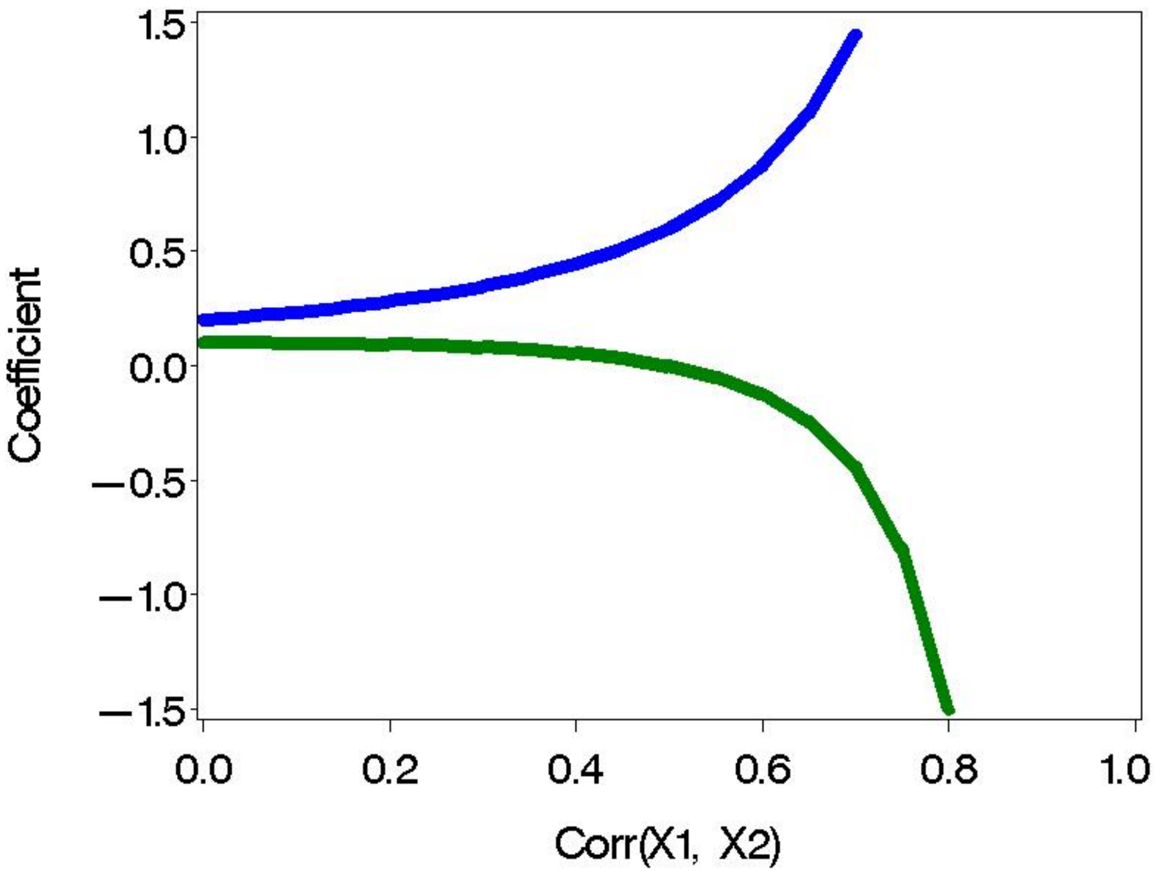


Illustration: Mitro et al, 2016 EHP

- **Background**: Exposure to persistent organic pollutants (POPs) such as **dioxins, furans, and polychlorinated biphenyls (PCBs)** may influence **leukocyte telomere length** (**LTL**), a biomarker associated with chronic disease.

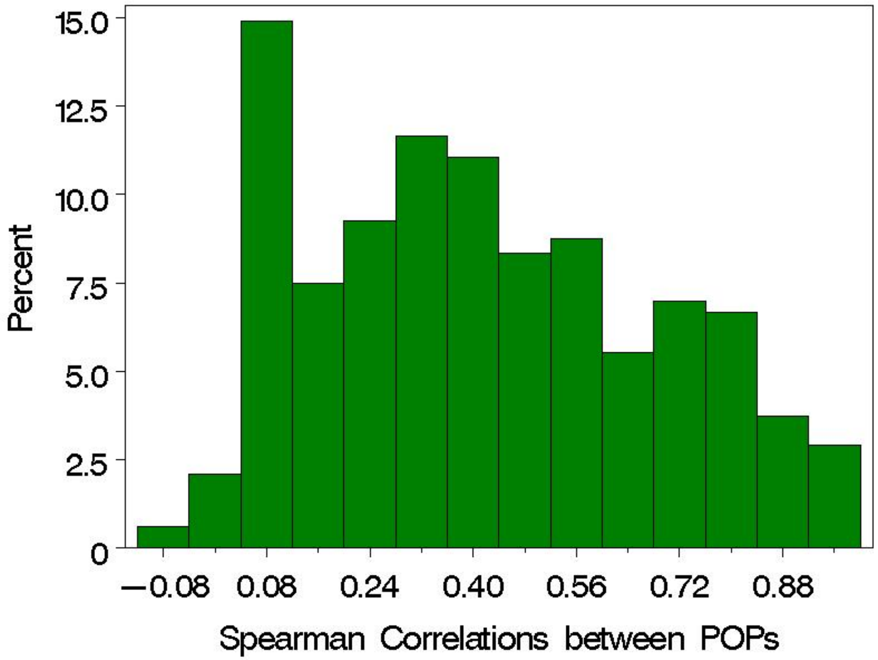
  In vitro research suggests dioxins may bind to the aryl hydrocarbon receptor (AhR) and induce telomerase activity, which elongates LTL.

  However, few epidemiologic studies have investigated associations between POPs and LTL.

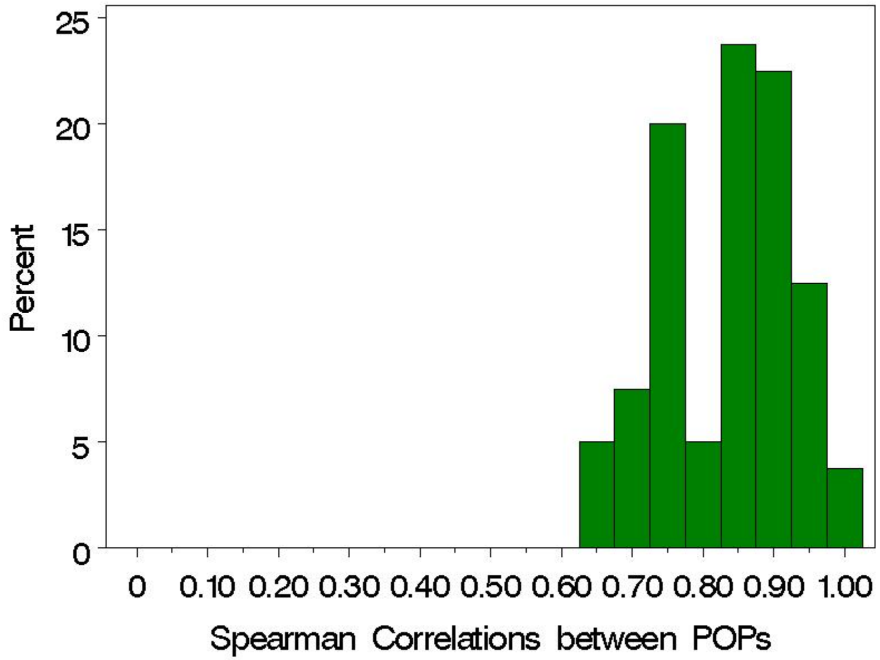
- **Covariates:**

  Models were adjusted for age, age2, sex, race/ethnicity, BMI, log(cotinine), white blood cell count, percent lymphocytes, percent monocytes, percent neutrophils, percent eosinophils, percent basophils

Correlation Between POPs



Full set of 18 POPs



Subset of 9 PCBs

STABILITY OF ILL-CONDITIONING WITH CONSTRAINTS: VARIANCE VS BIAS

Least Squares with Constraints

- Ridge Regression

$$\hat{\beta}_{ridge} = \min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

- LASSO

$$\hat{\beta}_{LASSO} = \min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right]$$

- Elastic Net

$$\hat{\beta}_{elastic\ net} = \min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left(\alpha \left| \beta_j \right| + (1-\alpha)\beta_j^2 \right) \right]$$

Weighted Quantile Sum (WQS) Regression (Carrico et al, 2014)

Nonlinear regression with weight parameters:

$$\boldsymbol{\theta} = \left[\beta_0, \beta_1, w_1, \ldots, w_c, \boldsymbol{\gamma}'\right]$$

$$g(\mu) = \beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \sum_{k=1} \gamma_k z_{ik}$$

Final WQS index is a weighted average across the bootstrap samples using a 'signal function'

$$WQS = \sum_{j=1}^{c} \bar{w}_j q_j$$

$$\bar{w}_j = \frac{1}{B} \sum_{b=1}^{B} w_{j(b)} f\left(\hat{\beta}_{1(b)}\right)$$

Final model:

$$g(\mu) = \beta_0 + \beta_1 WQS + \sum_{k=1} \gamma_k z_{ik}$$

Weighted Quantile Sum (WQS) Regression (Carrico et al, 2014)

Nonlinear regression with weight parameters:

$$\boldsymbol{\theta} = \left[\beta_0, \beta_1, w_1, \ldots, w_c, \boldsymbol{\gamma}'\right]$$

$$g(\mu) = \beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \sum_{k=1} \gamma_k z_{ik}$$

Why quantiles?

Final WQS index is a weighted average across the bootstrap samples using a 'signal function'

$$WQS = \sum_{j=1}^{c} \bar{w}_j q_j$$

$$\bar{w}_j = \frac{1}{B} \sum_{b=1}^{B} w_{j(b)} f\left(\hat{\beta}_{1(b)}\right)$$

Final model: $g(\mu) = \beta_0 + \beta_1 WQS + \sum_{k=1} \gamma_k z_{ik}$

Nonlinear Least Squares with Constraints

- WQS Regression with a Lagrange multiplier and with implicit directionality constraint

$$\hat{\theta}_{WQS} = \min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 \sum_{j=1}^{c} w_j q_j + \sum_{k=1} \gamma_k z_{ik} \right) \right)^2 + \lambda \left(\sum_{j=1}^{c} w_j - 1 \right) \right]$$

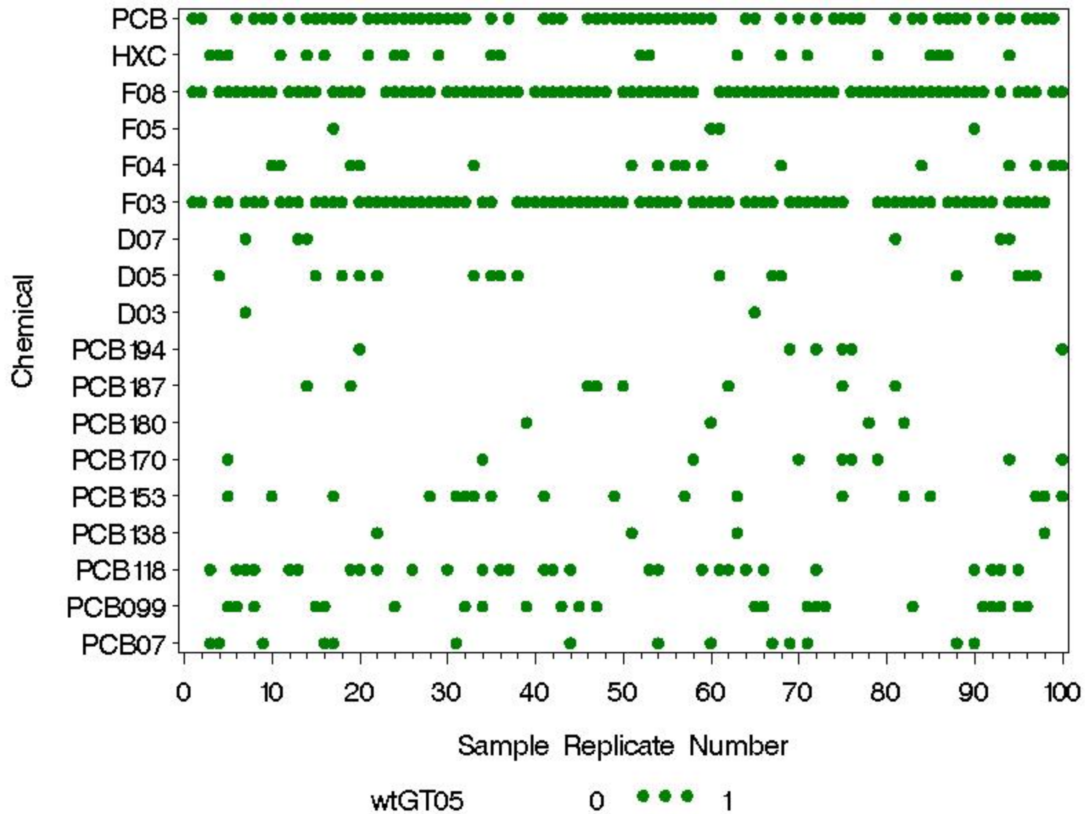
WQS regression: Ensemble step

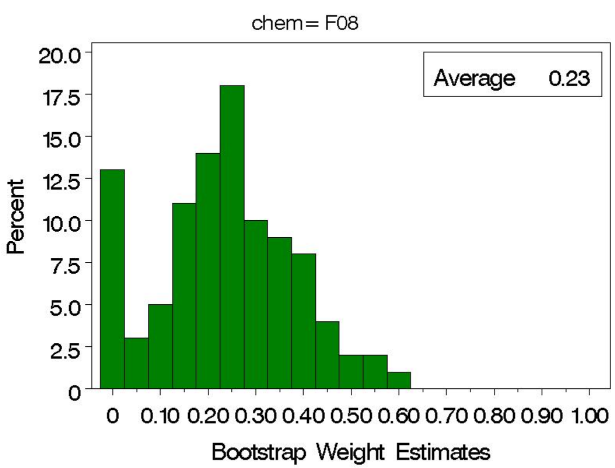
- Bootstrap samples of ***observations***
  - Why?
  - How many samples?
  - Distribution of weights

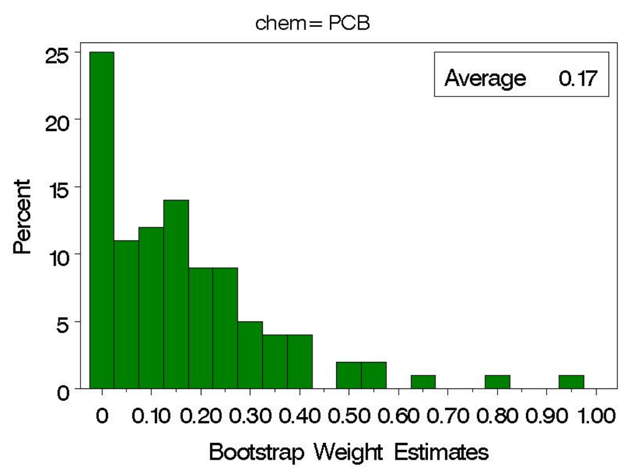
Two Strategies

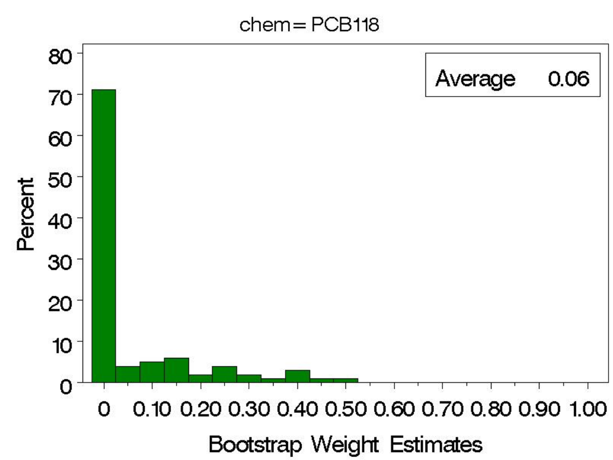
- Random subset of ***components*** (i.e., c variables)
  - Subsets of size, say, sqrt(c)
  - 1000 random subsets
  - Average across full set

Distribution of Weights across Bootstrap Samples









Splitting data for Training & Testing

- Generally, we use 40% of the sample for estimating weights and 60% for testing significance of the index
- Need more power for testing for significance of βετα 1

EXAMPLE: 9 PCBs and LTL

Preliminary adjusted analyses

Single chemical

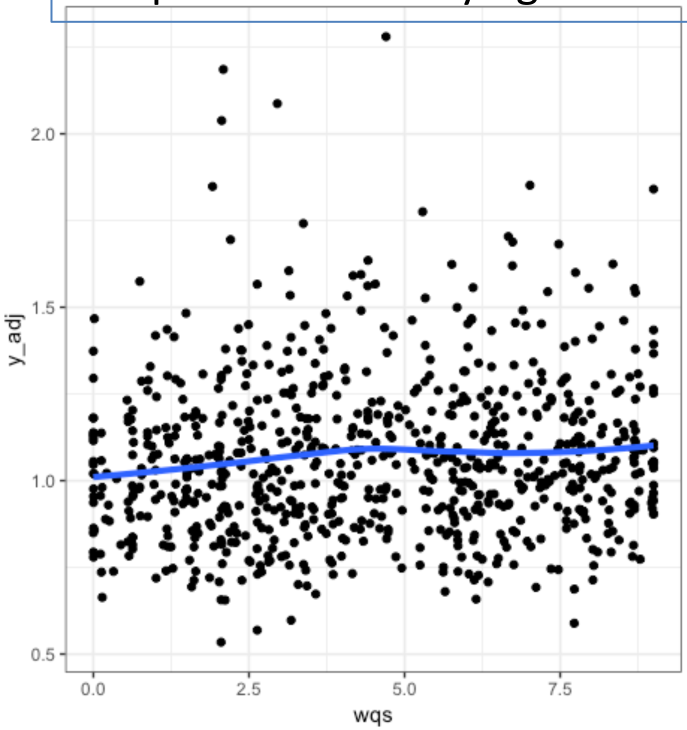
Parameter	Estimate	StdErr	ProbChiSq
log_LBX074LA	0.128	0.022	13E-9
log_LBX099LA	0.107	0.022	62E-8
log_LBX118LA	0.112	0.019	8E-9
log_LBX138LA	0.097	0.02	16E-7
log_LBX153LA	0.104	0.021	12E-7
log_LBX170LA	0.094	0.026	33E-5
log_LBX180LA	0.073	0.023	0.001
log_LBX187LA	0.085	0.024	46E-5
log_LBX194LA	0.061	0.028	0.032

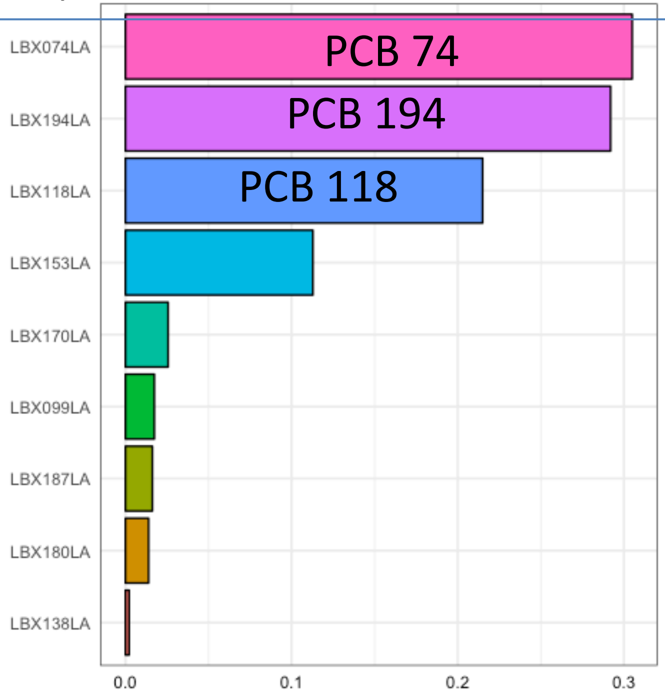
Joint model

Parameter	Estimate	Standard Error	Pr > ChiSq
logLBX074LA	0.0339	0.0197	0.0849
logLBX099LA	0.0037	0.0221	0.8674
logLBX118LA	0.0087	0.0193	0.6543
logLBX138LA	-0.0360	0.0354	0.3095
logLBX153LA	0.0904	0.0421	0.0315
logLBX170LA	-0.0015	0.0368	0.9664
logLBX180LA	-0.0348	0.0283	0.2181
logLBX187LA	-0.0077	0.0253	0.7603
logLBX194LA	-0.0019	0.0264	0.9423

EXAMPLE: WQS regression

Split: 40% for estimating weights; 60% for testing significance of WQS index
Quantiles: **deciles**
100 bootstrap samples
Analysis adjusted by covariates
Beta1 unconstrained
Cut-point for identifying a “bad actor”: 1/9 = 0.11

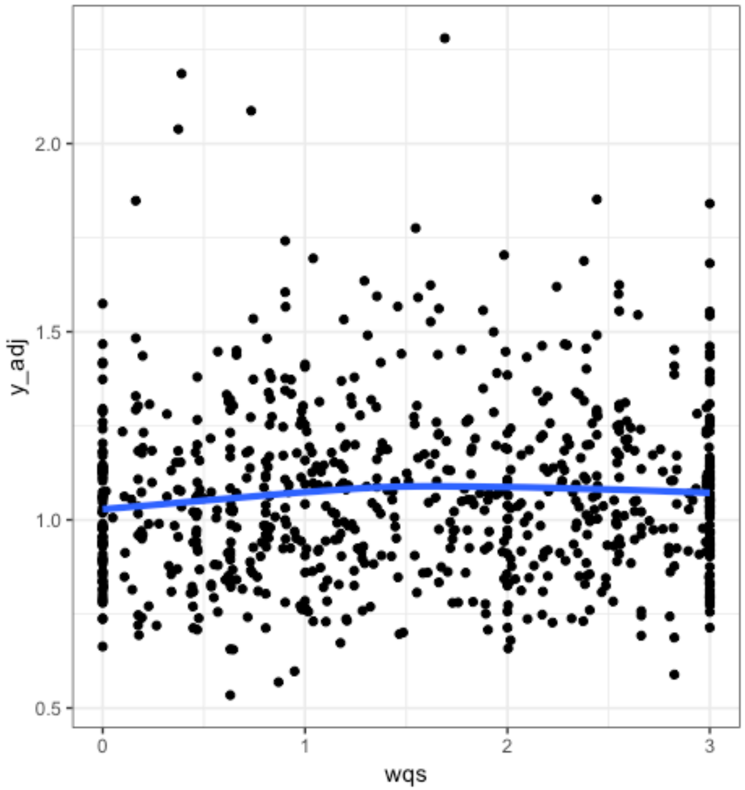


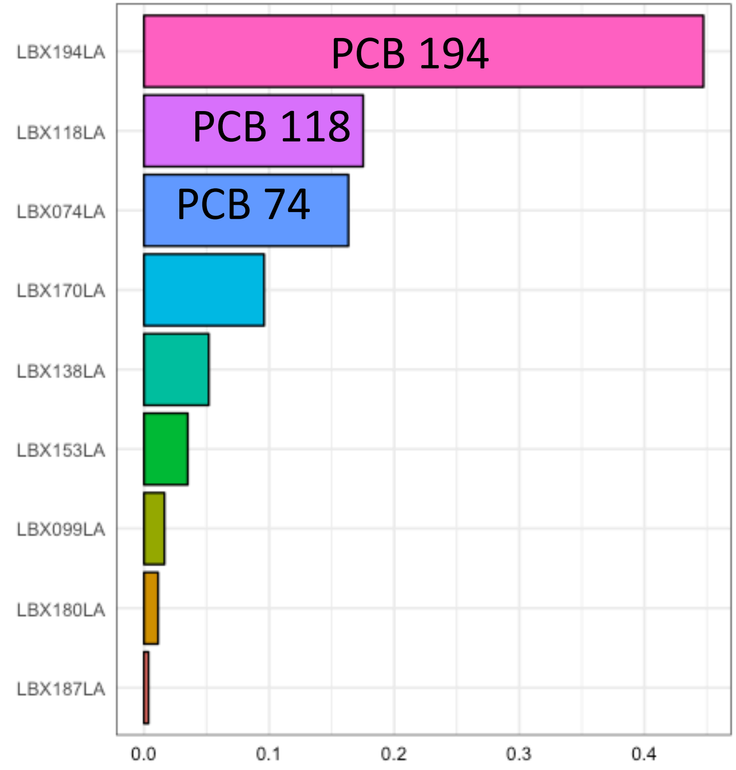


Beta1 = 0.023
SE= 0.005
p <0.001

EXAMPLE: WQS Regression

Split: 40% for estimating weights; 60% for testing significance of WQS index
Quantiles: **quartiles**
100 bootstrap samples
Analysis adjusted by covariates
Beta1 unconstrained
Cut-point for identifying “bad actor”: 1/9= 0.11





Beta1 = 0.042
SE= 0.014
P = 0.004

Stratified WQS regression

- Similar to interaction between a categorical variable and the weights
- Weights are estimated per each category in a single index where weights sum to 1
- STEPS (example: Brunst et al, 2017, AJE):
  - Determine overall quantiles per component
  - Use interaction quantile scoring; e.g.,

$$qscale\_white = \begin{Bmatrix} q, & \text{if white} \\ 0, & \text{otherwise} \end{Bmatrix}; \quad qscale\_nonwhite = \begin{Bmatrix} q, & \text{if nonwhite} \\ 0, & \text{otherwise} \end{Bmatrix}$$

RACE:	White		Nonwhite	
Stress scale	Weights	Cond Wt	Weights	Cond Wt
A	0.06	0.13	0.05	0.09
B	0.09	0.20	0.15	0.27
C	0.02	0.05	0.30	0.55
D	0.28	0.62	0.05	0.09
Sum	**45%**		**55%**	

Wrap-up

- Ill-conditioning due to multicollinearity in environmental health data is improved by constraints in the optimization for parameter estimation.
- Choice of strategy depends on the research question:
  - Biomarker identification (e.g., shrinkage methods)
  - Mixture effect (e.g., PCA, WQSR, BKMR)
  - Interaction among components (e.g., BKMR)

- **WQS regression** is based on quantile scores and is improved with the addition of the ensemble step
  - It addresses questions of a mixture effect with an empirically weighted index;
  - Stratified WQSR has the advantage that the sample size is not reduced to each strata
  - Extensions are forthcoming…

THANK YOU!