



VCU

Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations

Graduate School

2013

Characterization of a Weighted Quantile Score Approach for Highly Correlated Data in Risk Analysis Scenarios

Caroline Carrico
Virginia Commonwealth University

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>



Part of the [Biostatistics Commons](#)

© The Author

Downloaded from

<https://scholarscompass.vcu.edu/etd/3011>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

Copyright ©2013 Caroline K Carrico

All rights reserved

Characterization of a Weighted Quantile Score Approach for Highly Correlated Data in Risk Analysis Scenarios

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

By

Caroline Kelly Carrico

B.S. Winona State University, Winona, MN, 2008

Director: Chris Gennings, Professor, Biostatistics

Virginia Commonwealth University

Richmond, Virginia

March, 2013

Acknowledgements

First, I would like to thank my advisor and friend, Dr Chris Gennings for guiding my research, providing me with opportunities to grow as a researcher, and for being someone to share stories and experiences with. The rest of my committee: Dr Leroy Thacker, Dr Roy Sabo, Dr Joseph Ritter and Dr Pam Factor-Litvak were all instrumental in my thesis, each providing input and suggestions at every stage of the development. There are several more members of the Biostatistics department I would like to thank. Dr David Wheeler and Dr Will Anderson also contributed to this work despite not being committee members- offering key pieces of advice and knowledge. And last but certainly not least, Yvonne Hargrove and Cindy Sabo were always willing to help with anything and always had a friendly smile on their faces.

I also have to thank my family. I have to thank my Dad for telling me to “Go on, GET!” the day I packed up my car and headed to Virginia and for his infinite advice (relevant or not) and for always being proud of me. I’d also like to thank each of my siblings. We all tried to be sure that we were always there for each other growing up and I know that as adults they have always been and always will be there for me. I am proud of each of them and it has been their examples of success and hard work that have motivated me. And I want to thank my dearest friend, Helen, with whom I share a very special and unique bond.

I’d like to thank my husband, Bob Carrico, for blazing the trail before me and being there by my side from day one when he fixed my flat tire and gave me directions around a new city. Without him, I certainly wouldn’t be the person I am today or where I am today. He has always been there reminding and teaching me to be confident in all that I do. In addition, I’d like to thank his family for always making me feel like one of their own and giving me a home away from home.

I also have to thank my classmates for all the good times and for being there to commiserate during the bad times. I especially have to thank Bob and Mercer for taking me under their wings as the older and wiser, offering advice both academic and extracurricular- especially for teaching me the fine art of movie hopping. I also have to thank Amber for being physically next to me all throughout my time at VCU- we certainly gained a special bond sharing an office/cubicle wall every year. In addition, I want to thank Emily, Adam, Sarah, Stephanie for their friendship.

My cycling teammates and friends also need to be thanked for helping me enjoy my time away from school. The friendships I have gained through Altius Cycling Team, VCU Cycling Club, and the friendly faces I see at races all year have really been instrumental in my happiness during my time at VCU. Patricia Kinser was my first riding partner and has become one of my closest friends. In the years I have known her, she has constantly displayed a strength that I admire on and off the bike. I also want to thank ACT, especially Matt, Tha Peahen, and Chris Woods for their support and sense of humor. And thanks to MaryAnne for fueling my competition and motivation.

Table of Contents

List of Tables and Figures	v
Abstract	ix
Chapter 1: Introduction and Motivation	1
1.1 Motivation	2
1.2 Current Methods	3
1.3 Prospectus	6
 Chapter 2: Theoretical Justifications Through Heuristic Argument and Simulations (written as manuscript)	8
Title Page	9
Abstract	10
Introduction	11
1.1 Motivation	11
1.2 Current Methods	12
 Methods	13
2.1 Model and General Method Steps	13
2.2 Heuristic Argument for Improved Stability	15
 Simulations.....	21
3.1 Simulating Correlated Data.....	21
3.2 Simulation Results: Single Estimation of Weights.....	24
3.2.1: Validity of Weights: Single Sample Estimation.....	25
3.2.2: Reliability of Weights: Single Sample Estimation.....	28
3.3 Per Sample Estimation Simulations Conclusions	30
3.4 Simulation results: Validity and Reliability Across Bootstrap Samples.....	31
3.5 Traditional Methods Comparison.....	41
3.5.1 Ordinary Regression and LASSO Simulations.....	41
3.5.2 Direct Comparison to LASSO.....	44
 Conclusions.....	46

Chapter 3: Appendix: Supplementary Material to Chapter 2.....	48
Introduction.....	49
3.1 Various Simulations for Weighted Quantile Score with Phthalate Correlation Structure from Figure 1.1 in Chapter 2.....	50
3.2 Simulated Bootstrap Analyses For Breakdown Cases with Increased Sample Size.....	56
3.3 Comparison to LASSO.....	60
3.4 Applying to Real Data: Demonstration of Development of WQS.....	67
Chapter 4: Application of Method: Environmental Chemicals, Non-Chemical Stressors and Liver Health (written as manuscript).....	72
Title Page	73
Abstract	74
Introduction.....	75
Methods	76
4.2.1 Description of Data.....	76
4.2.2 Preliminary Statistical Analysis	79
4.2.3 Weighted Quartile Scores	80
Results	82
4.3.1 Preliminary Results	82
4.3.2 Weighted Index for Environmental Chemicals Score (ECS)	86
4.3.3 Weighted Index for Nutrient Stressor Score (NSS)	87
4.3.4 Joint Model for Chemicals and Nutrients	89
Discussion	90
4.4.1 Implication of Indices	90
4.4.2 Limitations	93
Chapter 5: Conclusions and Future Work	95
5.1 Conclusions	96
5.2 Future Work	97
5.3 Conclusion	100
Appendix I: References	101
Appendix II: SAS Code	105
A2.1 Simulating Correlated Data	106
A2.2 Macro for Simulating Bootstrap Samples	107
A2.3 Chapter 4 Code for ECS/NSS, ALT Analysis	112

List of Tables and Figures

Chapter 1

Figure 1.1	Correlation Structure of Set of Eleven Phthalate Monoesters.....	1
Table 1.1	Percent Above LOD for 13 Phthalate Monoesters from NHANES 2007-2008....	3

Chapter 2

Figure 2.1	Correlation Structure (A) for MHH, MHP, MIB, MOH, MZP.....	15
Figure 2.2	Condition Number for Hessian from Weighted Quartile Score Approach..... (Hessian) and Multiple Regression ($X^T X$)	16
Table 2.1	Effects of a Specific Ridge Values on a Correlation Matrix.....	20
Figure 2.3	Schematic of Simulation Cases.....	21
Figure 2.4	Distributions for the Number of Components Assigned Weight to Assess..... Validity; All horizontal axes 0 to 11 and vertical 0 to 100.	23
Figure 2.5	Distribution of Weights Across Varying Pairwise Correlations and..... Correlations with Y; All Horizontal Axes from 0 to 1 and Vertical from 0 to 100.	26
Figure 2.6	Updated Schematic to Indicate Performance of WQS Estimation.....	28
Figure 2.7	Simulated Bootstrap Analyses: Distribution of Average Weights from 1000.... Simulated Bootstrap Analyses (i.e. Average weight from the 100 bootstrap samples from each of 1000 simulated datasets). Eight Components Correlated with Outcome at 0.1 level; X3, X7, X10 NOT correlated with outcome; Observed phthalate correlation structure	30
Figure 2.8	Four Cases (i.e. Four Corners) For Simulated Bootstrap Analysis.....	32
Figure 2.9	Distributions for Weights and Distribution of Power across: 1000 Simulated.... Datasets; 100 Bootstraps; Sample Size 250 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.1 Correlation with Y	35
Figure 2.10	Distributions for Weights and Distribution of Power across: 1000 Simulated.... Datasets; 100 Bootstraps; Sample Size 250 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.1 Correlation with Y; Pairwise Correlations Decreased by 43% (i.e. ridge 1.5)	36

Figure 2.11	Distributions for Weights and Distribution of Power across: 1000 Simulated....37 Datasets; 100 Bootstraps; Sample Size 250 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.3 Correlation with Y
Figure 2.12	Distributions for Weights and Distribution of Power across: 1000 Simulated....38 Datasets; 100 Bootstraps; Sample Size 250 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.3 Correlation with Y; Pairwise Correlations Decreased by 43% (i.e. ridge 1.5)
Figure 2.13	Ordinary Regression Simulation.....40
Figure 2.14	LASSO Simulation Results.....41
Figure 2.15	Number of Components Detected by LASSO Correctly and Incorrectly.....42

Chapter 3

Figure 3.1	MHH (X7) Correlated with Y (Corr=0.3), Sample Size 1000, Observed47 Phthalate Correlation Structure (Ch 2 Figure 1.1), No Other components Correlated with Outcome
Figure 3.2	MHH (X7), MOH (X10) Correlated with Y (Corr=0.3), Pairwise.....49 Correlation=0.92, Sample Size 1000, Horizontal Axis is 0 to 1 and Vertical Axis 0 to 100 for All Histograms
Figure 3.3	All Six High Molecular Weight Phthalates Correlated with Y of 0.5;.....51 Observed Phthalate Correlation (Figure 1.1 in Ch 2); Note: Highest pairwise correlations occur between high molecular weight phthalates; Final histogram is the distribution of the number of weights greater than 0.05; Components correlated with outcome were: X1, X2, X3, X7, X8, X10 and are indicated in the table with shaded background.
Figure 3.4	Distributions for Weights and Distribution of Power across: 1000 Simulated.....53 Datasets; 100 Bootstraps; Sample Size 500 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.1 Correlation with Y; No Ridge- Pairwise Correlations from Figure 2.1 in Ch 2.
Figure 3.5	Distributions for Weights and Distribution of Power across:.....55 1000 Simulated Datasets; 100 Bootstraps; Sample Size 500 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.1 Correlation with Y; Ridge 1.5- Pairwise Correlations from Figure XXX Reduced by 43%

Figure 3.6	LASSO Results: No Selection Criterion; LARS Algorithm.....	57
Figure 3.7	ADJRSQ Adjusted R-square statistic.....	58
Figure 3.8	AIC Akaike information criterion.....	59
Figure 3.9	AICC Corrected Akaike information criterion.....	60
Figure 3.10	CV Predicted residual sum of square with k -fold cross validation.....	61
Figure 3.11	PRESS Predicted residual sum of squares.....	62
Table 3.1	Results from Breakdown Case with Phthalates with Sample Size of..... 250 and 8 Phthalates Marked with Asterisk Correlated with Y at 0.1 Level	65
Figure 3.12	Distribution of Weights from 1000 Bootstrap Samples of size 250 from..... Data from Table 3.1	67

Chapter 4

Table 4.1a	Analytes Considered in Analyses.....	72
Table 4.1b	Dietary Nutrients Considered in Analyses	72
Table 4.2	Comparison of Test and Validation Datasets for Covariates and Outcome.....	76
Figure 4.1	Correlations Among Environmental Chemicals and Nutrients..... (NOTE: Only those significantly different from zero are displayed. 100/666(15%) were nonsignificant for Chemicals and 386/1596(24%) for nutrients	77
Figure 4.2	Distributions of Pairwise Correlations for Dioxin-Like and Non-Dioxin-Like....	78
Table 4.3	Core Model Assessment	79
Figure 4.3	Plot of Age vs ALT to Demonstrate Quadratic relationship.....	79
Table 4.4	Average Bootstrap Weights for Environmental Chemicals	80
Table 4.5	Model Results for Average Bootstrap ECS.....	81
Table 4.6	Nutrient Weights from Bootstrap Analysis	82
Table 4.7	Model Estimation with Average Bootstrap Weights for Nutrient..... Non-Chemical Stressors	83

Figure 4.4	Model Estimation With Chemical and Nutrient Index and Predicted.....84 Mean ALT for ECS and NSS at Average Levels for BMI, Gender, Race (binary), PIR, and Age
Figure 4.5	Contour Plots for Men and Women Average ALT Versus ECS and NSS.....86
Figure 4.6	Distribution of Correlations with Log(ALT) for Chemicals and Nutrients.....88

Chapter 5

Figure 5.1	Diagram of Possible Mediating Effect93
------------	--

Abstract

In risk evaluation, the effect of mixtures of environmental chemicals on a common adverse outcome is of interest. However, due to the high dimensionality and inherent correlations among chemicals that occur together, the traditional methods (e.g. ordinary or logistic regression) are unsuitable. We extend and characterize a weighted quantile score (WQS) approach to estimating an index for a set of highly correlated components. In the case with environmental chemicals, we use the WQS to identify “bad actors” and estimate body burden. The accuracy of the WQS was evaluated through extensive simulation studies in terms of validity (ability of the WQS to select the correct components) and reliability (the variability of the estimated weights across bootstrap samples). The WQS demonstrated high validity and reliability in scenarios with relatively high correlations with an outcome and moderate breakdown in cases where the correlation with the outcome was relatively small compared to the pairwise correlations. In cases where components are independent, weights can be interpreted as association with the outcome relative to the other components. In cases with complex correlation patterns, weights are influenced by both importance with the outcome and the correlation structure. The WQS also showed improvements over ordinary regression and LASSO in the simulations performed. To conclude, an application of this method on the association between environmental chemicals, nutrition and liver toxicity, as measured by ALT (alanine aminotransferase) is presented. The application identifies environmental chemicals (PCBs, dioxins, furans and heavy metals) that are associated with an increase in ALT and a set of nutrients that are identified as non-chemical stressors due to an association with an increase in ALT.

I. Introduction and Motivation

1.1 Motivation

In risk evaluation, the effect of mixtures of environmental chemicals on a common adverse outcome is of interest. However, due to the high dimensionality and inherent correlations among chemicals that occur together, the traditional methods (e.g. ordinary or logistic regression) are unsuitable. To illustrate, suppose we are interested in modeling risk for log HDL (high density lipoprotein, an indicator of high cholesterol and a factor associated with increased risk of cardiovascular disease) based on biomonitoring data for urinary levels of a set of phthalate monoesters. In this case, the correlation structure among the set of phthalate monoesters (these phthalate monoesters are explained further in Table 1.1) is given in Figure 1.1, using biomonitoring data from the National Health And Nutrition Examination Survey from the 2007-2008 cycle (CDC: NCHS, 2012).

Figure 1.1: Correlation Structure of Set of Eleven Phthalate Monoesters

	CNP	COP	ECP	MBP	MC1	MEP	MHH	MHP	MIB	MOH	MZP
CNP	1	0.40	0.24	0.06	0.37	0.03	0.17	0.16	0.05	0.19	0.04
COP		1	0.41	0.13	0.56	<0.01	0.34	0.20	0.17	0.37	0.13
ECP			1	0.23	0.39	0.04	0.84	0.53	0.20	0.85	0.18
MBP				1	0.34	0.15	0.28	0.19	0.51	0.29	0.39
MC1					1	0.01	0.36	0.21	0.21	0.40	0.25
MEP						1	0.05	0.04	0.14	0.05	0.03
MHH							1	0.59	0.28	0.92	0.24
MHP								1	0.18	0.57	0.09
MIB									1	0.30	0.28
MOH										1	0.23
MZP											1

Note: Data collected from spot urine, adjusted for creatinine, and categorized into quartiles (0-3)
 Bold values= significant at 0.05 level

The correlation structure for these chemicals is complex, ranging from near 0 correlations to near perfect correlation (0.92). Due to the complex nature of the correlations, traditional regression models would suffer from problems with variance inflation of parameter estimates.

1.2 Current Methods

Several techniques have been proposed to combat this problem, including Ridge Regression, LASSO, and the Elastic Net. In ridge regression, all p predictors remain in the model but are biased slightly to decrease the variance of the parameter estimates. This prevents a parsimonious model and complicates the interpretability (Zou and Hastie 2005). The LASSO technique was developed by Tibshirani (1996) to improve accuracy (by reducing parameter estimate variance) and allow for better interpretability. The LASSO method imposes a tuning parameter on the parameters which forces some variables to zero while others are minimized until the residual sums of squares is minimized and the sum of the absolute value of the parameters is less than a specified constant (Tibshirani, 1996). The elastic net, like the LASSO, shrinks variance and selects a subset of the original predictors through a regularization bias on the original predictors (Zou and Hastie 2005). While these methods are convenient and well-supported, they have limitations. As stated before, ridge regression models do not reduce the dimensionality of the problem. In the presence of high correlations among predictor variables, the LASSO method has been shown to select an arbitrary member from the group (Zou and Hastie, 2005). The elastic net method has a “grouping effect” that causes correlated predictors to either all be eliminated from the model or all used in the model (Zou and Hastie, 2005).

These methods are more suitable if prediction is the primary purpose of the research. However, when the goal of the model is to evaluate relationship or determine risk of given predictors on an outcome, there is reason to consider an alternative.

A common method for determining risk between highly correlated environmental chemicals and a health outcome is to consider only the single chemical, single outcome effect.

This method may not be affected by variance inflation but it does not take into account the mixture effect of environmental chemical exposures. Biomonitoring data from NHANES shows that exposure to multiple environmental chemicals is widespread. Subjects in the NHANES dataset have both blood serum levels and urinary levels of an extensive number of chemicals measured. A given subject has levels above the limit of detection (LOD) on most chemicals evaluated. Consider the 1732 subjects with the correlation structure from Figure 1.1, the percent above LOD for each of these phthalate monoesters is given in Table 1.1.

Table 1.1: Percent Above LOD for 13 Phthalate Monoesters from NHANES 2007-2008

Phthalate Monoesters (Abbreviation)	% Above LOD
Mono(carboxynonyl) Phthalate (CNP)	90
Mono(carboxyoctyl) Phthalate (COP)	96
Mono-2-ethyl-5-carboxypentyl phthalate (ECP)	100
Mono-n-butyl phthalate (MBP)	99
Mono-(3-carboxypropyl) phthalate (MC1)	98
Mono-ethyl phthalate (MEP)	100
Mono-(2-ethyl-5-hydroxyhexyl) phthalate (MHH)	100
Mono-(2-ethyl)-hexyl phthalate (MHP)	67
Mono-isobutyl phthalate (MIB)	100
Mono-(2-ethyl-5-oxohexyl) phthalate (MOH)	98
Mono-benzyl phthalate (MZP)	98

For the eleven phthalates given in Table 1.1, the limit of detection was at least 65% for these eleven phthalates. There were four other phthalates measured in the NHANES subsample; the percent above LOD for those four was less than 50%. Because these chemicals are detectable in such a high percentage of subjects (in a very large, national sample) and because they are so highly correlated, analyzing their effect on a health outcome simultaneously with current methods (ordinary regression) or individually does not take into account the possible mixture

effect. There is a need for a method that takes into account the complexity of the exposure pattern, but is more robust to multicollinearity.

Swan, et al (2008) proposed using a score created by quartile scoring phthalate monoester levels and then adding the total exposure amount. By quartile scoring, the effect of differing potencies and the skewness in exposure patterns are controlled. This method also will not suffer from the multicollinearity issues in ordinary regression. What this method lacks is interpretability. In a risk analysis setting, the goal is to detect “bad actors,” which has motivated our method.

We propose extending the work of Gennings, et al (2010) and Christensen, et al (2013) by using a weighted index in which weights are empirically determined and are calculated to optimize the likelihood of the desired model. The components of the index are selected based on logical groupings and the weights are constrained to sum to 1 and be between 0 and 1 allowing them to be interpreted as an index for body burden. The weights are estimated from the data using a bootstrap analysis with validation in an independent validation dataset. This method reduces the dimensionality and the issues with multicollinearity while maintaining interpretability. We define and characterize this approach in terms of the validity and reliability of the weights. The validity of the weights is determined by the WQS approach’s ability to place weights on the correct components. The reliability of the weights is assessed by the variance of the assigned weights. Both are evaluated through simulation.

1.3 Prospectus

The goal of this thesis is to extend, characterize, and apply the weighted quantile score approach. Chapter 2 presents a heuristic argument for the increased stability of the proposed method over ordinary regression and LASSO. Then, through extensive simulations, we evaluate the validity (ability to detect components that are simulated to be “bad actors”) and the reliability of the weights estimated (i.e. the variability of the estimated weights across the bootstrap samples). We use validity and reliability to compare the method at hand to LASSO.

Chapter 3 contains supplementary material for Chapter 2. This material includes additional simulation cases: varying sample sizes, different correlation patterns (correlation with outcome changed and/or pairwise correlations altered). A comparison to LASSO from Chapter 2 is also extended to include different selection criterion. The extra simulations are assessed for their validity and reliability, especially as compared to LASSO and the simulations given in Chapter 2. The chapter concludes with a real data example of the improvement associated with the bootstrap analysis.

Chapter 4 is an application of the method using NHANES data and modeling liver toxicity. In this chapter, we not only estimate an index for a set of environmental chemicals, but also for a non-chemical stressor on liver health. In risk assesment focus was shifted to such “non-chemical stressors” after the National Research Council’s report in 2009 recommended the consideration of both chemical and non-chemical stressors on public health (Lewis, 2011). Common non-chemical stressors include socio-economic status, race/ethnicity, obesity, occupational and community-related exposures, and many more. We estimate a nutritional index to determine if poor nutrition is a non-chemical stressor for liver health. Other non-chemical

stressors like gender, age, race/ethnicity, BMI, and poverty:income ratio are considered as covariates, but not in an index.

Chapters 2 and 4 in this thesis are written as standalone manuscripts for submission for peer review publication, so there may be repeated information.

The overall goal of this thesis is to extend and characterize the weighted quantile score approach for highly correlated data in a risk analysis setting. We show that theoretically, the approach has improved stability due to the addition of the constraint for the optimization. We demonstrate this improved stability along with improved false positive and false negative rates through simulations. We show that the weighted quantile score approach may outperform both ordinary least squares and LASSO methods.

II. Characterization of a Weighted Quantile Score Approach for Highly Correlated Data in a Risk Analysis Setting

Introduction

1.1 Motivation

In risk evaluation, the effect of mixtures of environmental chemicals on a common adverse outcome is of interest. However, due to the high dimensionality and inherent correlations among chemicals that occur together, the traditional methods (e.g. ordinary or logistic regression) suffer from multicollinearity and variance inflation. To illustrate, suppose we are interested in modeling risk for log HDL (high density lipoprotein, an indicator of poor cardiovascular health) based on biomonitoring data for urinary levels of a set of phthalate monoesters. In this case, the correlation structure among the set of phthalate monoesters is given in Figure 1.1, using biomonitoring data from the National Health And Nutrition Examination Survey from the 2007-2008 cycle (CDC: NCHS, 2012). The correlation structure for these chemicals is complex, ranging from near 0 correlations to near perfect correlation (0.92). Due to the complex nature of the correlations, traditional regression models would suffer from problems with variance inflation of parameter estimates.

1.2 Current Methods

Several techniques have been proposed to combat this problem, including Ridge Regression, LASSO, and the Elastic Net. In ridge regression, all p predictors remain in the model but are biased slightly to decrease the variance of the parameter estimates. This prevents a parsimonious model and complicates the interpretability (Zou and Hastie 2005). The LASSO

technique was developed by Tibshirani (1996) to improve accuracy (by reducing parameter estimate variance) and allow for better interpretability. The LASSO method imposes a tuning parameter on the parameters which forces some variables to zero while others are minimized until the residual sums of squares is minimized and the sum of the absolute value of the parameters is less than a specified constant (Tibshirani, 1996). The elastic net, like the LASSO, shrinks variance and selects a subset of the original predictors through a regularization bias on the original predictors (Zou and Hastie 2005). While these methods are convenient and well-supported, they have limitations. As stated before, ridge regression models do not reduce the dimensionality of the problem. In the presence of high correlations among predictor variables, the LASSO method has been shown to select an arbitrary member from the group (Zou and Hastie, 2005). The elastic net method has a “grouping effect” that causes correlated predictors to either all be eliminated from the model or all be used in the model (Zou and Hastie, 2005).

These methods are more suitable if prediction is the primary purpose of the research. For example, if the goal is to show association between phthalates as a whole and HDL, LASSO, ridge regression, or elastic net may be suitable. But when the objective is to determine which phthalates in particular are associated with HDL, a different method is needed. This has motivated our proposed method, the weighted quantile score approach.

Extending and characterizing the work of Gennings, et al (2010) and Christensen, et al (2013), we propose using a weighted linear index in which weights are empirically determined through bootstrap sampling. The components of the index are selected based on logical groupings of components that occur together and would have a common adverse outcome. The weights are constrained to sum to 1 and be between 0 and 1, reducing the dimensionality and the issues with multicollinearity while maintaining interpretability. We define and characterize this

approach in terms of the validity and reliability of the weights. The validity of the weights is determined by the WQS approach's ability to place weights on the correct components (in simulations, the correlation with the outcome will be set, so the correct components are known). The reliability of the weights is assessed by the variance of the assigned weights. Both are evaluated through simulation.

Methods

2.1 Model and General Method Steps

Consider data with correlated components (c components) that are reasonable to combine into an index. Let the values for the c components be scored into quartiles, denoted q_i for $i=1$ to c . The data (total sample size $N=N_1+N_2$) are first split into a test (N_1) and a validation dataset (N_2). Bootstrap samples of size N_1 are generated from the test dataset (typically $B=1000$ bootstrap samples) and are used to estimate the unknown weights, w_i , that maximize the likelihood for the model for $b=1$ to B :

$$g(\mu)=\beta_0+\beta_1 * \sum_{i=1}^c w_i * q_i + \mathbf{z}^T \boldsymbol{\phi} \Big|_b$$

Where w is a $c \times 1$ vector of weights, w_i for the i^{th} component q_i (2.1)

$$\text{with } \sum_{i=1}^c w_i \Big|_b = 1 \text{ and } 0 \leq w_i \leq 1 \Big|_b$$

Using the above notation, g represents any monotonic, differentiable link function as in a generalized linear model, which links the mean, μ , to the predictor variables. The term, $\sum_{i=1}^c w_i * q_i$ represents the weighted index for the set of c chemicals of interest, and w_i represents the weight associated with the i^{th} components (whose quantile score is denoted q_i). The

covariates of interest are accounted for in the vector, \mathbf{z} and need to be determined prior to estimating the weights. In order to accomplish weights that are empirically based, each bootstrap sample is used to estimate the weights that maximize the likelihood for Equation 2.1. These estimated weights are tested in each bootstrap sample. The weights that are significantly validated in each bootstrap sample are used to estimate the weighted quantile score, WQS:

$$g(\mu) = \beta_0 + \beta_1 * WQS + \mathbf{z}^T \boldsymbol{\phi}$$

$$\text{where } WQS = \sum_{i=1}^c \bar{w}_i * q_i \quad (2.2)$$

$$\bar{w}_i = \frac{1}{n_B} \sum_{i=1}^{n_B} w_i, \quad n_B = \text{number of bootstrap samples in which } \beta_1 \text{ was significant}$$

In order to empirically and simultaneously estimate the weights and the parameters, we employ optimization algorithms that maximize a continuous nonlinear function subject to a linear constraint, $\sum_{i=1}^c w_i = 1$ and bounds $w_i \in [0,1]$. So, for our case, we have a general nonlinear optimization function subject to one linear constraint (boundaries are not effected by optimization methods as they just limit the parameter space). Optimization algorithms available include, Trust Region Method, Newton-Raphson with Line Search or Ridging, the Quasi-Newton Method, and the Conjugate Gradient Methods (SAS 9.2 Documentation). We have chosen Trust Region method for this paper, because it allows for a linear constraint on a nonlinear objective function and was stable. A description of this optimization strategy is given in Numerical Optimization by Nocedal and Wright (1999). The NLP procedure in SAS 9.2 treats the constrained optimization in the Lagrange format under the Kuhn-Tucker Conditions (SAS Manual).

2.2 Heuristic Argument for Improved Stability

Common uses of constrained optimizations include ridge regression and the LASSO models. The forms of these models (Hastie, et al 2009) are:

$$\begin{aligned}\hat{\beta}_{\text{ridge}} &= \arg \min_{\beta} \left[\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^{p-1} x_{ij} * \beta_j)^2 + \lambda (\sum_{j=1}^p \beta_j^2 - t) \right] \\ \hat{\beta}_{\text{lasso}} &= \arg \min_{\beta} \left[\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^{p-1} x_{ij} * \beta_j)^2 + \lambda (\sum_{j=1}^p |\beta_j| - t) \right]\end{aligned}\tag{2.3}$$

Both constrained optimization problems are of the Lagrangian form with different constraints. The parameter, λ , is the Lagrange multiplier and t is a presepecified tuning parameter. Hoerl and Kennard (1970) demonstrated that in a regression case with multicollinearity, the added constraint in a ridge regression setting stabilizes the estimate $\hat{\beta}_{\text{ridge}}$ by making the parameter space “more orthogonal” (i.e. decreasing the effect of multicollinearity). Hoerl and Kennard demonstrate that the added constraint results in a smaller range in the eigenvalues (i.e. a smaller eigenvalue spectrum). Similarly, the proposed weighted quantile score model in (2.1) can be written in the Lagrangian format. In this framework, for $g(\mu) = \mu$, the form of the model (2.1) would be:

In least squares optimization, for a parameter vector

$$\theta = [\beta_0 \quad \beta_1 \quad w_1 \quad w_2 \quad \cdots \quad w_{c-1} \quad \phi]:$$

$$\hat{\theta}_{\text{wqs}} = \arg \min_{\theta} \left[\sum_{i=1}^n (y_i - \beta_0 + \beta_1 * \sum_{i=1}^c w_i * q_i + \mathbf{z} \cdot \phi)^2 + \lambda (\sum_{i=1}^c w_i - 1) \right]\tag{2.4}$$

or, equivalently, in maximum likelihood form:

$$\hat{\theta}_{\text{wqs}} = \arg \max_{\theta} [\ln(L(X, \theta)) - \lambda (\sum_{i=1}^c w_i - 1)]$$

Under this form of the equation, for each bootstrap sample, the log-likelihood for the model in equation 2.4 is optimized subject to the constraint that the sum of the weights is equal to 1. Following the argument from Hoerl and Kennard, the constraint stabilizes the estimation process by reducing the eigenvalue spectrum.

A proxy for measuring this increased stability as indicated through the eigenvalue spectrum is through the condition number of a matrix, which is defined as the ratio of the largest to the smallest eigenvalue. We followed the methods of Anderson (2008) to evaluate the stability of the weighted quantile score approach compared to ordinary regression.

Consider the normal equations for the estimation for an ordinary regression model:

$$\begin{aligned}(X'X)\beta &= X'Y \\ \hat{\beta} &= (X'X)^{-1} X'Y\end{aligned}$$

In an ordinary regression model, the estimation step involves taking the inverse of $X'X$.

In order to estimate the parameters in the weighted quantile score approach, we need to derive the optimization steps. In order to incorporate the linear constraint into the optimization,

we define the weights such that $w_c = 1 - \sum_{i=1}^{c-1} w_i$ and therefore, the vector of parameters, θ , is

defined as:

$$\theta' = [\beta_0 \quad \beta_1 \quad w_1 \quad w_2 \quad \cdots \quad w_{c-1} \quad \phi]$$

Using a Taylor series expansion for estimation in a nonlinear model we have the following:

$$\begin{aligned}\frac{\delta\mu}{\delta\theta} &= G(\theta) = G(\theta^s) + \frac{\delta G}{\delta\theta}(\theta^{s+1} - \theta^s) \\ \text{note: } \frac{\delta G}{\delta\theta} &= \frac{\delta\mu}{\delta\theta} * \frac{\delta\mu}{\delta\theta} = -\frac{\delta\mu}{\delta\theta\delta\theta} = -H(\theta) \\ \text{Setting equal to zero to optimize:} \\ 0 &= G(\theta^s) + \frac{\delta G}{\delta\theta}(\theta^{s+1} - \theta^s) \\ 0 &= G(\theta^s) - H(\theta^s) * (\theta^{s+1} - \theta^s) \\ H(\theta^s) * (\theta^{s+1} - \theta^s) &= G(\theta^s) \\ (\theta^{s+1} - \theta^s) &= H^{-1}(\theta^s) * G(\theta^s) \\ \theta^{s+1} &= \theta^s + H^{-1}(\theta^s) * G(\theta^s)\end{aligned}$$

Where θ^{s+1} denotes the updated parameter estimates, θ^s denotes the estimates from the current step, H denotes the hessian matrix and G , the gradient vector both evaluated at the current step parameter estimates, θ^s . So in the weighted quantile score approach, the optimization is contingent on the stability of $H^{-1}(\theta)$. Through derivation, the form of $H(\theta)$ is:

$$H(\theta) = \begin{bmatrix} 2n & 2\sum_{i=1}^n [\sum_{j=1}^{c-1} (w_j * x_{ji}) + (1 - \sum_{k=1}^{c-1} w_k) * x_{ci}] & 2\beta_1 \sum_{i=1}^n x_{1i} - x_{ci} & 2\beta_1 \sum_{i=1}^n x_{2i} - x_{ci} & \cdots & 2\beta_1 \sum_{i=1}^n x_{c-1,i} - x_{ci} \\ \sum_{i=1}^n [\sum_{j=1}^{c-1} (w_j * x_{ji}) + (1 - \sum_{k=1}^{c-1} w_k) * x_{ci}]^2 & 2\beta_1 \sum_{i=1}^n (x_{1i} - x_{ci})^2 & 2\beta_1 \sum_{i=1}^n (x_{2i} - x_{ci})^2 & \cdots & 2\beta_1 \sum_{i=1}^n (x_{c-1,i} - x_{ci})^2 \\ & 2\beta_1 \sum_{i=1}^n (x_{1i} - x_{ci})^2 & 2\beta_1 \sum_{i=1}^n (x_{1i} - x_{ci}) * (x_{2i} - x_{ci}) & \cdots & 2\beta_1 \sum_{i=1}^n (x_{1i} - x_{ci}) * (x_{c-1,i} - x_{ci}) \\ & & 2\beta_1 \sum_{i=1}^n (x_{2i} - x_{ci})^2 & \cdots & 2\beta_1 \sum_{i=1}^n (x_{2i} - x_{ci}) * (x_{c-1,i} - x_{ci}) \\ & \text{SYM} & & \ddots & \vdots \\ & & & & 2\beta_1 \sum_{i=1}^n (x_{c-1,i} - x_{ci})^2 \end{bmatrix}$$

Based on the results from Hoerl and Kennard, we propose the precision of our method is attributed to the increased stability of $H^{-1}(\theta)$ over $X^T X$. Both the hat matrix and the Hessian matrix were column-scaled (i.e. each element was divided by the norm of the column vector) to have unit length in order to allow for comparison of the condition numbers. We then calculated

the condition number, $\kappa(\mathbf{A})$, as an indication of the stability of a matrix and the increased uniformity of the eigenvalue spectrum, using definition that the condition number is equal to the ratio of the maximum and minimum singular values:

$$\kappa(\mathbf{A}) = \frac{\sigma_{max}}{\sigma_{min}}$$

The singular values of a matrix \mathbf{A} are defined as the values σ_i such that $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$, where $\mathbf{\Lambda}$ is a diagonal matrix with the value σ_i on the i^{th} diagonal. Here \mathbf{U} is composed of the left-singular vectors of \mathbf{A} (i.e. eigenvectors of $\mathbf{A}\mathbf{A}^*$) and \mathbf{V} contains the right-singular eigenvectors of \mathbf{A} (i.e. eigenvectors of $\mathbf{A}^*\mathbf{A}$). In the case of a square matrix, the singular values are the absolute values of the eigenvalues of the matrix. So in the case of a square matrix, the condition number is equal to:

$$\kappa(\mathbf{A}) = \frac{|\lambda_{max}|}{|\lambda_{min}|}$$

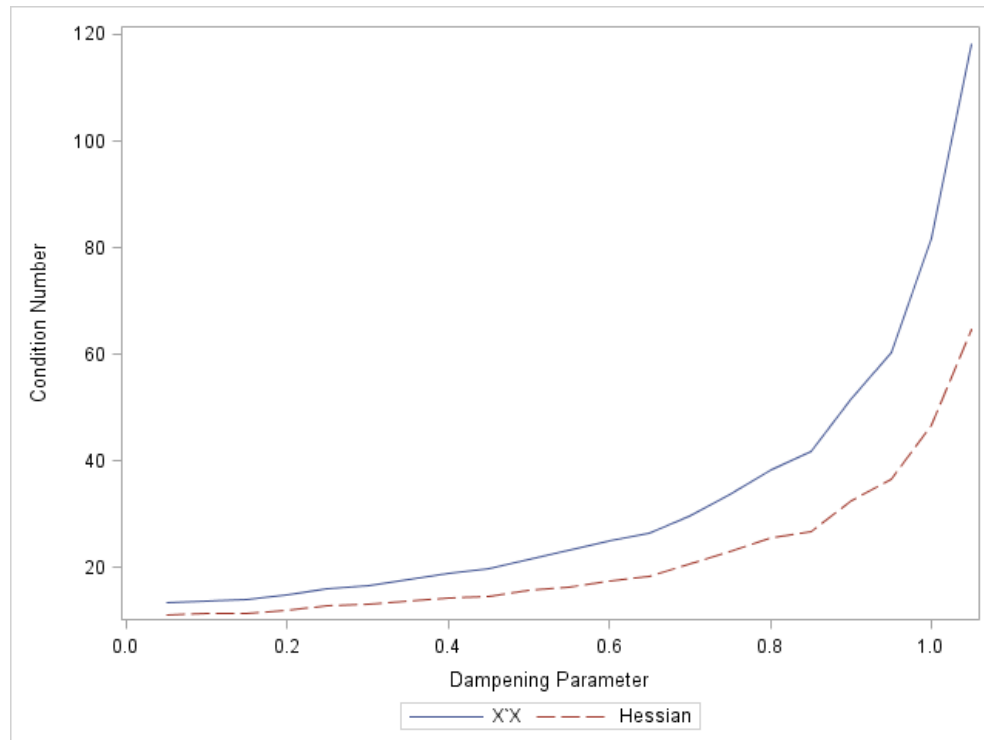
To investigate the stability of both $\mathbf{X}^*\mathbf{X}$ and $\mathbf{H}(\boldsymbol{\theta})$, we simulated data for a set of 5 phthalates (MHH, MHP, MIB, MOH AND MZP) with the following observed correlation structure in Figure 2.1. Details for simulating correlated data are given in Section 3.1.

Figure 2.1: Correlation Structure (\mathbf{A}) for MHH, MHP, MIB, MOH, MZP

	1	0.59	0.28	0.92	0.24
		1	0.18	0.57	0.09
			1	0.3	0.28
				1	0.23
					1

We used dampening parameters (m) from 0.05 to 1.05 (by 0.05) to see the tendencies as the correlation increases from 5% of the above correlations to 105% of the above correlations, using the equation $\mathbf{A}^* = (\mathbf{A} - \mathbf{I})m + \mathbf{I}$ for $m = 0.05$ to 1.05 by 0.05. As in Hoerl and Kennard with ridge regression, we found that the parameter space is more orthogonal since the condition number is always smaller and increases much less severely than that of a multiple regression. This suggests that the estimates from the weighted quantile score approach are more stable and precise than those from ordinary regression models. Results are given in Figure 2.2.

Figure 2.2: Condition Number for Hessian from Weighted Quartile Score Approach (Hessian) and Multiple Regression ($X^T X$)



Simulations

3.1 Simulating Correlated Data

Our objective is to simulate normally distributed data $N(M, \Sigma)$ with a given correlation structure for an outcome y and predictors x_1, x_2, \dots, x_c . Let:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \text{corr}(y, x_1) & \text{corr}(y, x_2) & \cdots & \text{corr}(y, x_c) \\ & 1 & \text{corr}(x_1, x_2) & \cdots & \text{corr}(x_1, x_c) \\ & & 1 & \ddots & \vdots \\ & & & \ddots & \text{corr}(x_{c-1}, x_c) \\ & & & & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{Var}(y) & \text{Cov}(y, x_1) & \text{Cov}(y, x_2) & \cdots & \text{Cov}(y, x_c) \\ & \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \cdots & \text{Cov}(x_1, x_c) \\ & & \text{Var}(x_2) & \vdots & \vdots \\ & & & \ddots & \text{Cov}(x_{c-1}, x_c) \\ & & & & \text{Var}(x_c) \end{bmatrix}$$

$$\mathbf{m} = \begin{bmatrix} \bar{y} \\ \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_c \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} \sqrt{\text{Var}(y)} \\ \sqrt{\text{Var}(x_1)} \\ \vdots \\ \sqrt{\text{Var}(x_c)} \end{bmatrix} = \begin{bmatrix} \text{SD}(y) \\ \text{SD}(x_1) \\ \vdots \\ \text{SD}(x_c) \end{bmatrix}$$

To impose the correlation structure, we first use the relationship between the correlation and the variance that yields:

$$\boldsymbol{\Sigma} = \text{diag}(\mathbf{S}) * \boldsymbol{\rho} * \text{diag}(\mathbf{S})$$

Then follow the following simulation steps (let $p=c+1$):

- 1) Calculate the Cholesky Decomposition of Σ ($p \times p$ dimension), such that $\Sigma = \mathbf{U}_{p \times p} \mathbf{U}_{p \times p}^T$. (For full detail on this calculation, see Harville, 2008)
- 2) Simulate $\mathbf{Z}_i \sim N(\mathbf{0}_{p \times 1}, \mathbf{I}_p)$. $\mathbf{Z} = [\mathbf{Z}_1 \mathbf{Z}_2 \dots \mathbf{Z}_n]$, so \mathbf{Z} will have $n \times p$ dimension where each observation is generated from a standard normal distribution.
- 3) Let $\mathbf{M} = (\mathbf{m}^* \mathbf{1}_{1 \times n})^T$, $\mathbf{Y}_{n \times p} = \mathbf{M}_{n \times p} + \mathbf{Z}_{n \times p} * \mathbf{U}_{p \times p}$ and the i^{th} row is $\mathbf{Y}_i = \mathbf{m}_{p \times 1} + \mathbf{U}_{(p \times p)}^T * \mathbf{Z}_i (p \times 1)$
 - a) $E(\mathbf{Y}) = E(\mathbf{M} + \mathbf{Z} * \mathbf{U}) = \mathbf{M} + E(\mathbf{Z}) = \mathbf{M}$
 - b) $\text{Var}(\mathbf{Y}_i) = \text{Var}(\mathbf{m} + \mathbf{U}^T * \mathbf{Z}_i) = \text{Var}(\mathbf{m}) + \text{Var}(\mathbf{U}^T * \mathbf{Z}_i) = \mathbf{0} + \mathbf{U}^T \mathbf{U} = \Sigma$
- 4) So, \mathbf{Y}_i is $N_p(\mathbf{m}, \Sigma)$

In the first step, in order to calculate \mathbf{U} , Σ must be positive definite. To evaluate relevant cases with highly correlated data, Σ may be near singular. In this case, we use matrix ridging to stabilize the matrix. Ridging a matrix involves adding a constant to the values on the diagonal of the matrix. If Σ is not positive definite:

- Define r , a ridge value (see table 3.1 for indication of the effect a given r will have on the correlations)

$$\bullet \quad \rho^* = \begin{bmatrix} 1+r & \text{corr}(y, x_1) & \text{corr}(y, x_2) & \dots & \text{corr}(y, x_c) \\ & 1+r & \text{corr}(x_1, x_2) & \dots & \text{corr}(x_1, x_c) \\ & & 1+r & \ddots & \vdots \\ & & & \ddots & \text{corr}(x_{c-1}, x_c) \\ & & & & 1+r \end{bmatrix}$$

- $\Sigma^* = \text{diag}(S) * \rho^* * \text{diag}(S)$
- Follow steps 1-4 with the above substitutions

The higher the ridge, the greater the impact on the values of the matrix. Table 2.1 lists the average multiplier for a given element of the matrix for ridge values from 0 to 0.5. So, for example, adding a ridge of 0.1 to the diagonals (i.e. making the diagonals of the correlation matrix 1.1) causes an average reduction in the correlation matrix of 20%. So to simulate data with a correlation with Y of 0.3, the input correlation is 0.375. Using a ridge of 0.5 has a reduction of 43% on the correlations, so the input correlation (to achieve a correlation of 0.3) is 0.526. So, when a large ridge is used, the pairwise correlations are reduced but a higher correlation with the outcome can still be simulated. Using a ridge in this situation is appropriate when the resulting simulated correlations are stated as target and not the correlations before the ridge is applied. Using the ridging, we can also see how the method performs as the correlation with Y becomes greater than the pairwise correlations among the components of the index.

Table 2.1: Effects of a Specific Ridge Values on a Correlation Matrix

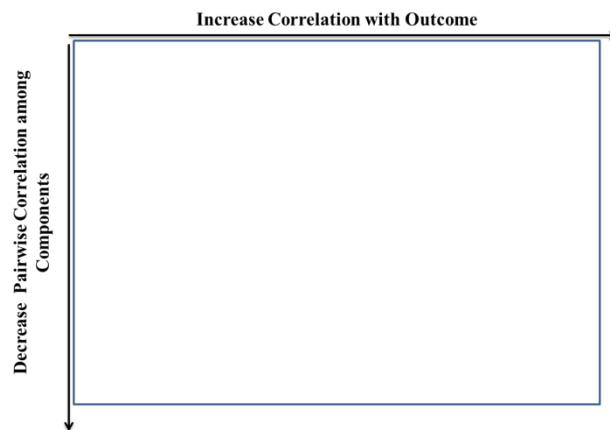
Diagonal Value (1+r)	Effective multiplier on Correlation matrix	Target Correlation with Y	Input Correlation with Y
1	1	0.1	0.100
		0.2	0.200
		0.3	0.300
1.05	0.82	0.1	0.122
		0.2	0.244
		0.3	0.366
		0.4	0.488
1.1	0.8	0.1	0.125
		0.2	0.250
		0.3	0.375
		0.4	0.500
1.2	0.72	0.1	0.139
		0.2	0.278
		0.3	0.417
		0.4	0.556
1.3	0.67	0.1	0.149
		0.2	0.299
		0.3	0.448
		0.4	0.597
1.4	0.61	0.1	0.164
		0.2	0.328
		0.3	0.492
		0.4	0.656
1.5	0.57	0.1	0.175
		0.2	0.351
		0.3	0.526
		0.4	0.702
		0.5	0.877

3.2 Simulation Results: Single Estimation

To characterize the weighting procedure, we need to verify two things. First, validity: i.e., the number of components assigned weights is appropriate (i.e. the weighting procedure is picking up all the important factors). Validity will be defined for a given cutpoint; that is a component will be deemed “selected” if its weight is greater than a chosen cutpoint. Second,

reliability: i.e., the weights that are assigned are reliable (determined by the variation in the weight estimates). We used simulation studies to assess the performance of the weighted quartile score method in terms of these two aspects of accuracy, across varying levels of correlation with the outcome and degrees of multicollinearity (i.e. pairwise correlation structures), as shown in the schematic in Figure 2.3:

Figure 2.3: Schematic of Simulation Cases



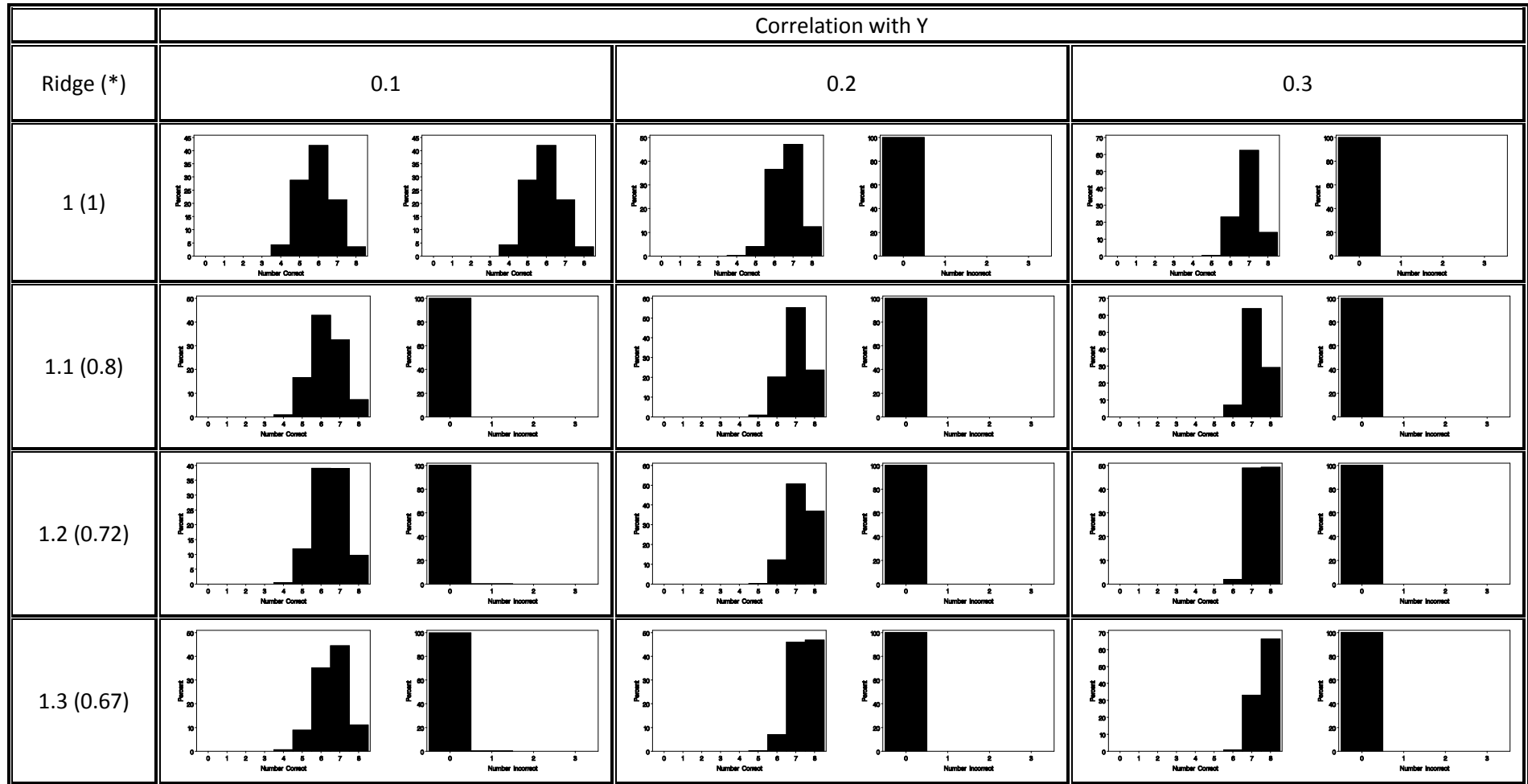
In the simulations, we based our data on biomonitoring data on phthalate levels in adult subjects from the 2007-2008 NHANES cycle. The correlation matrix for the quartile-scored phthalate monoesters is given in Figure 1.1. We simulated 1000 studies each with a sample size of 2000 observations. We simulated an outcome variable based on the observed distribution of HDL in the same population from which the phthalates correlation structure was derived.

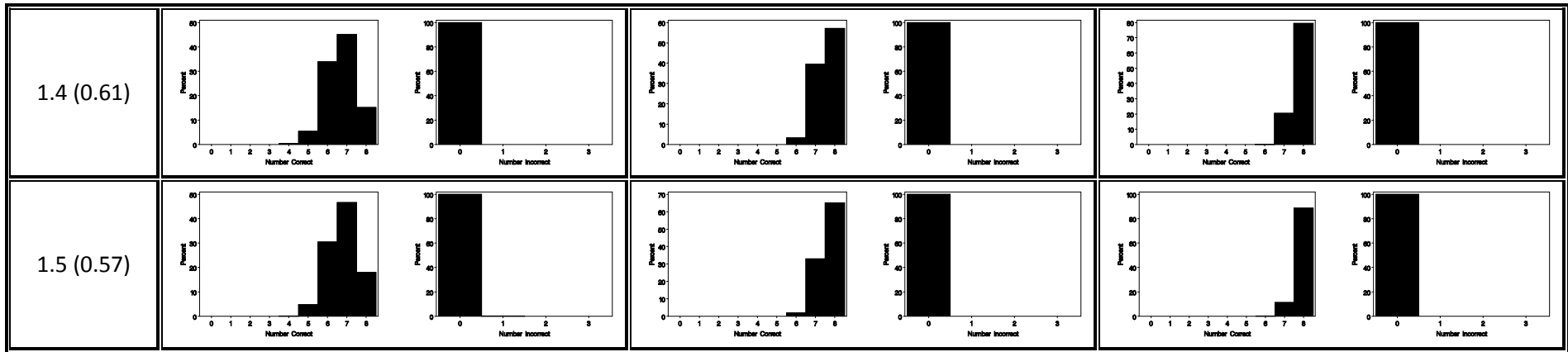
3.2.1: Validity of Weights: Single Sample Estimation

To determine how well the weighted quartile score performs in terms of validity, we began by determining to what extent the weighted quartile score detects important factors. We

chose eight of the eleven phthalates to have correlation patterns that cover the schematic in Figure 2.3 by varying the correlation between the eight components (namely X1, X2, X4, X5, X6, X8, X9, X11) and the outcome and among the eleven components by using different ridge values. We defined validity as the number of correct components assigned a weight of at least 0.05. The results, given in Figure 2.4, indicate that as the pairwise correlations among the phthalate monoesters decreases, validity increases. That is, for a given correlation with the outcome, as the pairwise correlations decrease, number of components assigned weight tends towards the truth (assigning weight to eight of the eleven components). The same is true for the correlation with Y: as the correlation with Y increases, the method has higher validity. With low correlation with Y and no decrease in the pairwise correlations, the distribution number of components assigned weight is centered at 6.5 and ranges from 4-8. As the correlation and the ridge increase, the center of the distribution shifts to 8 and the range tightens. This demonstrates that the method performs well when the correlation with the outcome is relatively large compared to the pairwise correlations and worse when the correlation with Y is relatively small compared to the pairwise correlations. It is promising that in the worst case considered, the method is still able to detect all but one of the important components on average. For the definition of validity, a cutoff value must be chosen. Here, a cutoff of 0.05 was used, but it is likely that for a higher number of components and/or a more complex correlation structure, a smaller cutoff value may need to be used. The number of components in the weighted index should be considered when determining a cutpoint. To help guide the cutpoint, consider the average weight if all components are assigned a weight (i.e. for 20 components the average weight would be 0.05, so if there are greater than 20 components a smaller cutpoint should be used).

Figure 2.4: Distributions for the Number of Components Assigned Weight to Assess Validity; All horizontal axes for Correct 0 to 8 and 0 to 3 for Incorrect; Sample size for test dataset: 1000



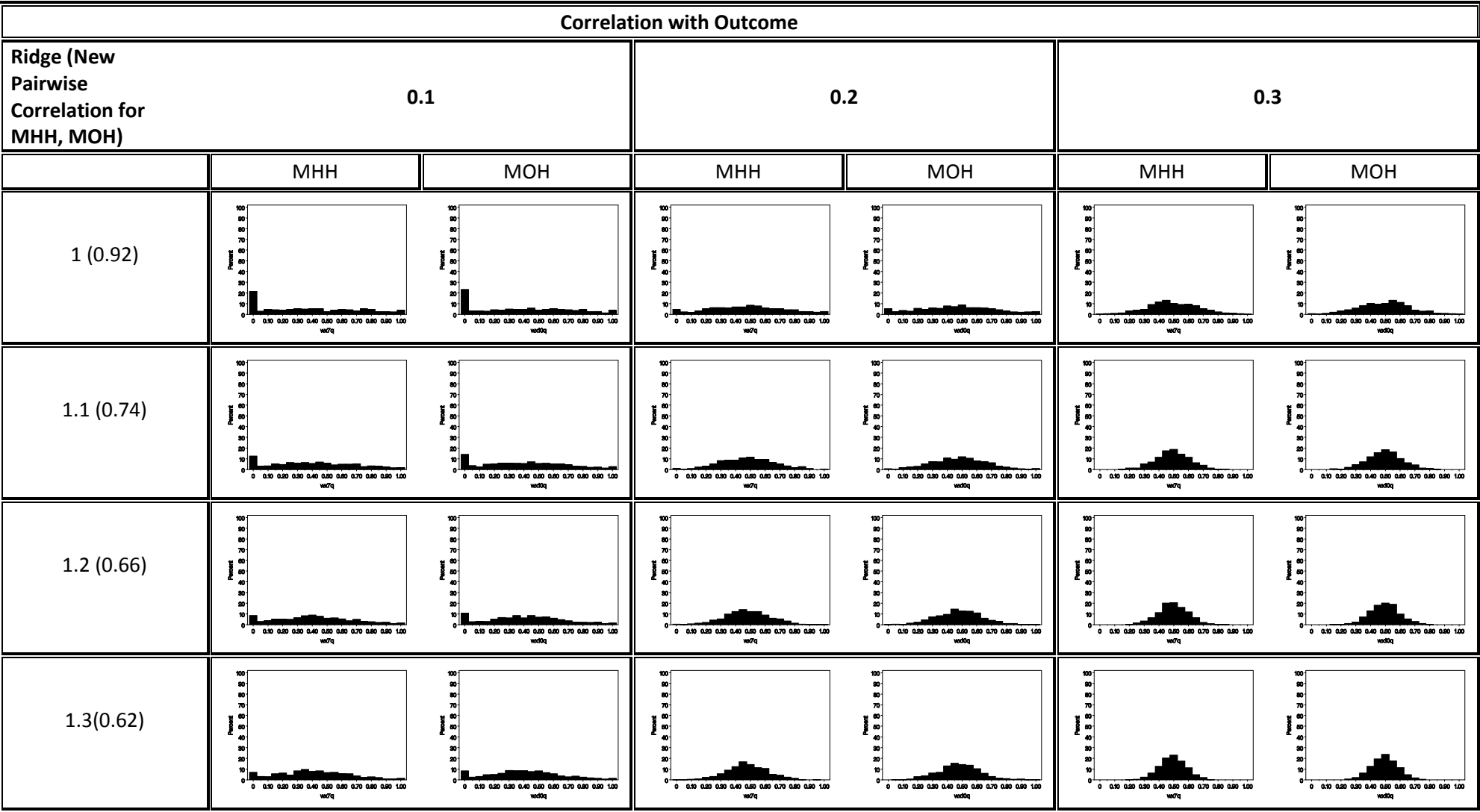


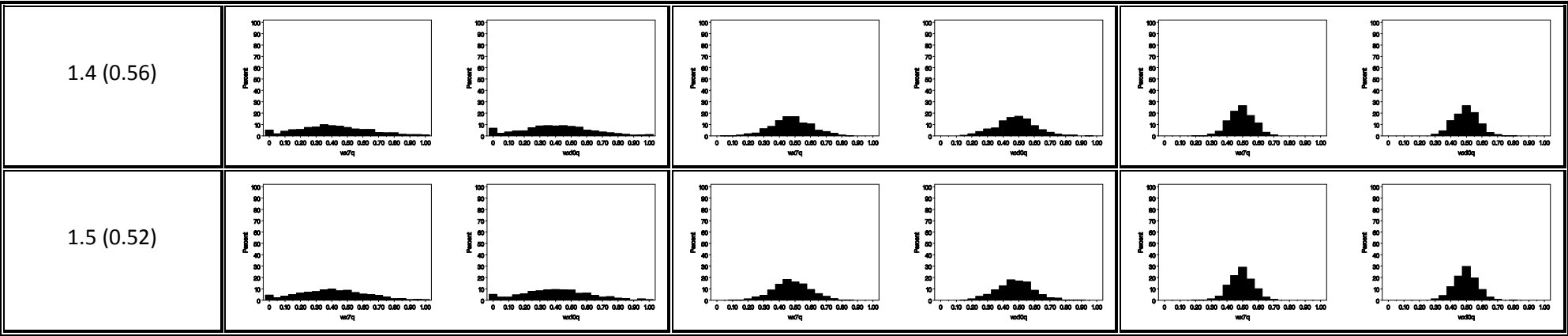
*Resulting Multiplier on Correlations

3.2.2: Reliability of the Weights: Single Sample Estimation

After determining that the weighted quartile score estimates weights on an appropriate number of components, it was important to determine the reliability of the weights assigned to the components. For that reason, we considered two phthalates with high pairwise correlations (MOH and MHH, correlation is 0.92). We again used a correlation with Y ranging from 0.1 to 0.3 and a ridge of 0 to .5 (i.e. pairwise correlation ranges from 0.52 to 0.92) to cover a portion of the schematic in Figure 2.3. Shown in Figure 2.5, we found again that the method is improved as the correlation between the important factors and the outcome variable becomes relatively larger compared to the pairwise correlations. The weights are centered about 0.50 (equal weight on the two components) and as the correlations change (i.e. as the correlation with Y increases and the pairwise correlation decreases), the reliability increases (i.e. the variance of the distribution of the weights becomes small).

Figure 2.5: Distribution of Weights Across Varying Pairwise Correlations and Correlations with Y; All Horizontal Axes from 0 to 1 and Vertical from 0 to 100.

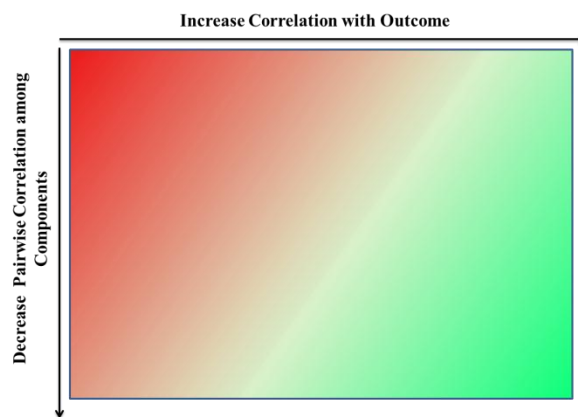




3.3 Per Sample Estimation Simulation Conclusions

In a risk analysis setting, it is important to detect the components that are associated with the outcome and to exclude those that are unrelated to the outcome. From the simulations, the weighted quartile score typically did not place weight on components that are unrelated to the outcome (indicated by weight greater than 0.05). The performance of the WQS depended on the setting. The results indicate that the correlation between the important components and the outcome has a greater effect on the performance of the estimation than the pairwise correlations among the components. The schematic from 2.3 has been updated to demonstrate where the method performs well (green) and where it tends to break down (red). Because the pairwise correlations among components cannot be altered, it is ideal that they will have less of an effect on the stability of the method. The correlation with the outcome can be altered by selecting an outcome that has a strong (i.e. not trivial) relationship to the components.

Figure 2.6: Updated Schematic to Indicate Performance of WQS Estimation



3.4: Simulation Results: Validity and Reliability Across Bootstrap Samples

To further evaluate validity and reliability, we performed simulation studies wherein we simulated 1000 datasets of sample size 250 with the phthalate correlation structure and eight of the eleven phthalates correlated with outcome at varying levels and observed pairwise correlations. The WQS index was calculated for each of the 1000 datasets (using the 100 bootstrap samples for each). For the WQS, validity is based on whether or not the average weight for a component (that is correlated with the outcome) is greater than a specified amount (depending on number of components and complexity of the correlation pattern). A cutoff of 0.05 was used to determine whether or not a component had been selected.

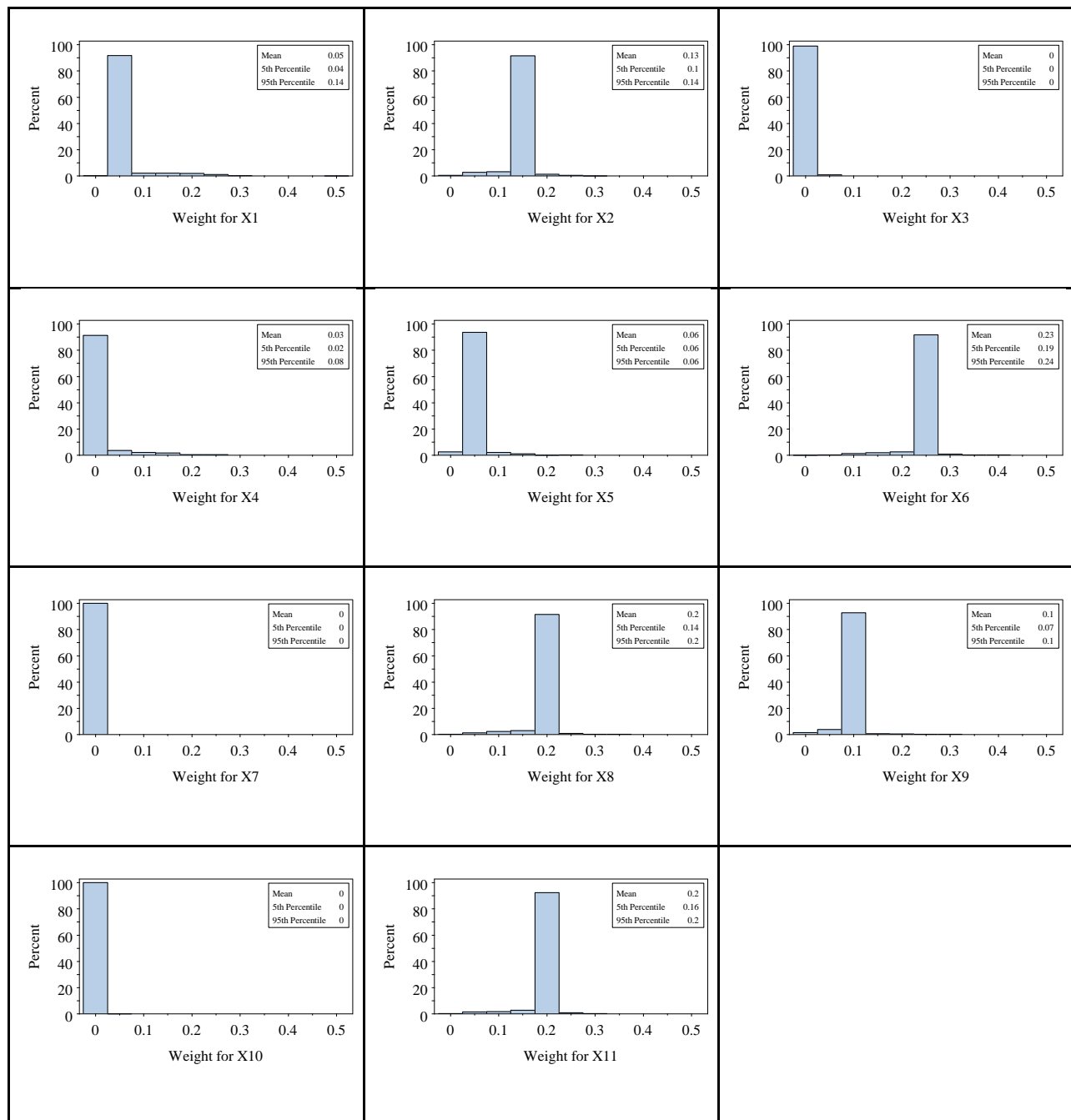
The first setting used, allowed for eight of the eleven components to be correlated with the outcome, namely X1, X2, X4, X5, X6, X8, X9, X11. By selecting these components, the highest pairwise correlations were excluded, effectively diminishing the effects of the multicollinearity. Among the phthalates, X3, X7 and X10 have the highest pairwise correlations (ranging from 0.82 to 0.92). The remaining eight phthalates have correlations typically less than 0.50. The results of this simulation study are given in Figure 2.7. Each histogram represents the distribution of the average weight with the average, 5th-percentile and the 95th-percentile given in the inset.

Figure 2.7: Simulated Bootstrap Analyses: Distribution of Average Weights from 1000

Simulated Bootstrap Analyses (i.e. Average weight from the 100 bootstrap samples from each of

1000 simulated datasets)- 8 Components Correlated with Outcome at 0.1 level; X3, X7, X10

NOT correlated with outcome; Observed phthalate correlation structure; sample size 250

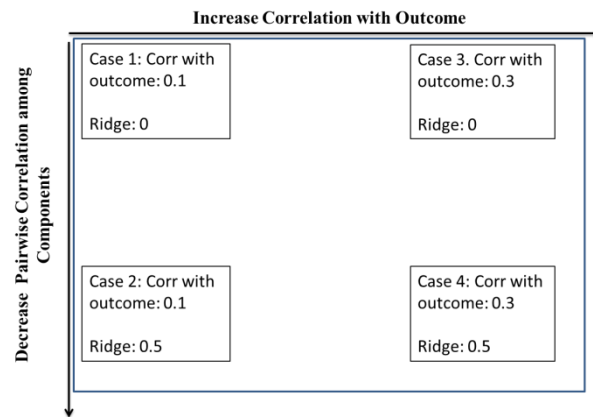


The results in 2.7 indicate both the validity and reliability are high. The average weight for the three components not correlated with the outcome was 0 for all and the weights for components selected as correlated with the outcome were all, on average, greater than 0. The reliability was also high as indicated by the lack of variation in the average weights (indicated by the “tower-looking” distributions). While the method was able to distinguish the important components from those that are not associated with the outcome, the pairwise correlations still had an impact on the estimation of the weights. Without any knowledge of the results, one would have likely predicted that the eight components would have equal weights. Because the average weight differs across the eight components, there is indication that something is affecting the estimation of the weights, likely the pairwise correlations among the eight components. From the simulated bootstraps, we see that across all the simulations, the components with the lowest average of the bootstrap sample average weights were CNP(X1), MBP(X4), MC1(X5), and MIB(X9). Each of these had an average bootstrap sample average of less than 0.1 (Note with equal weights on the 8 components, expected weight would be 0.125). For these four components, we suspect that the correlations with the remaining four components are relatively high and therefore affecting the distribution of the weights.

Since there is an indication of the effect of pairwise correlations on both the single estimation step and the bootstrap analysis, further simulations were performed to determine the extent of the effect of the pairwise correlations on the weighted quantile score approach. For the further analyses, we considered four corners from the schematic in Figure 2.3 (shown in Figure 2.8). We randomly selected three of the eleven components to be unrelated to the outcome. Then we altered the correlation with the outcome (between 0.1 and 0.3) and used either no ridge or a

ridge of 0.5, which imposed a decrease in the pairwise correlations of 0.43 (i.e. a multiplier of 0.57 on each of the pairwise correlations, see Table 1.1). This created the four new simulation cases (i.e. the four corners). The three components that were randomly chosen to have no correlation with the outcome were CNP, COP and MEP.

Figure 2.8: Four Cases (i.e. Four Corners) For Simulated Bootstrap Analysis



The results for these simulations are given in Figures 2.9-2.12 on the following pages. The final distribution in each of the figures is the distribution of the power (i.e. the proportion of bootstrap samples whose weights are validated in the validation dataset) across the 1000 simulated datasets. In the first case, Figure 2.9 (original pairwise correlations and correlation with the outcome of 0.1), the results indicate that there is little distinction between the components that should be assigned weight and several of the components that were correlated with the outcome (i.e. decreased validity) and that the distributions have more variation than those in Figure 2.10 (i.e. decreased reliability). However, the power distribution shows that these poor results may just be an indication of the lack of power and lack of information in the data at hand. In the second case, where the correlation with the outcome is again 0.1 but the pairwise correlations are ridged, we see minimal improvement in validity, reliability, and power (Figure 2.10). In the third case (Figure 2.11), where the correlation with the outcome for the eight

components is increased to 0.3, there is a great increase in validity, reliability and the power. Six of the eight components have weight distributions clearly different from zero and two of the three that should be zero have distributions almost always at 0. The fourth case (Figure 2.12), demonstrated little improvement in validity in case 3, but there is improvement in the reliability over case 3 as the distributions seemed to tighten (i.e. decrease variability and increase reliability).

It is promising to see that the increase in the correlation with the outcome is the change that is needed to increase the reliability of the weights along with the power of the method. The decrease in the pairwise correlations doesn't seem to have a strong effect on either the reliability of the weights or the power. This shows that the method will perform well despite high pairwise correlations among the predictors as long as there is a relatively strong relationship with the outcome.

These simulations also provide more information about the trends in the approximate weights. All simulation cases had equal correlation among all the components that were set to be correlated with the outcome, but there were differences between the average weights. The simulations suggest that these differences could be explained by the pairwise correlations. That is, components with high pairwise correlations seem to have diminished weights, despite having the same pairwise correlation with the outcome. For example, ECP (X3), MHH (X7), MOH (X10) have high pairwise correlations (0.82-0.92) and have the lowest weights of the eight that are simulated to be correlated with the outcome. This is most apparent in Figure 2.12 where the reliability in the weights and the power of the analyses are highest. This result means that in order to interpret weights, pairwise correlations need to be considered. As the number of components and the complexity of the pairwise correlations increase, this will become less and

less feasible. In such cases, the index may be interpreted as a whole, rather than individual weights.

Figure 2.9: Distributions for Weights and Distribution of Power across: 1000 Simulated Datasets; 100 Bootstraps; Sample Size 250 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.1 Correlation with Y

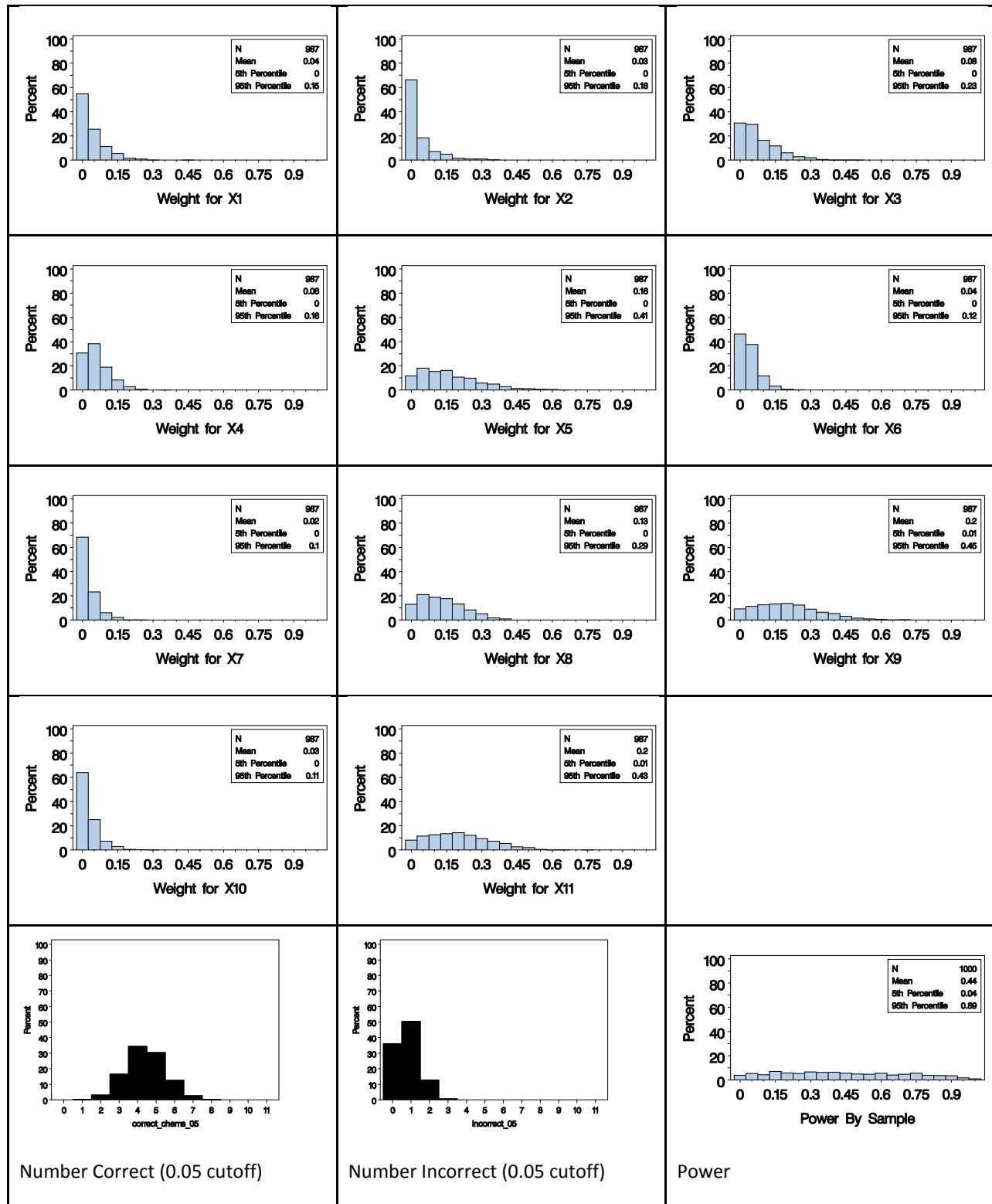


Figure 2.10: Distributions for Weights and Distribution of Power across: 1000 Simulated Datasets; 100 Bootstraps; Sample Size 250 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.1 Correlation with Y; Pairwise Correlations Decreased by 43% (i.e. ridge 1.5)

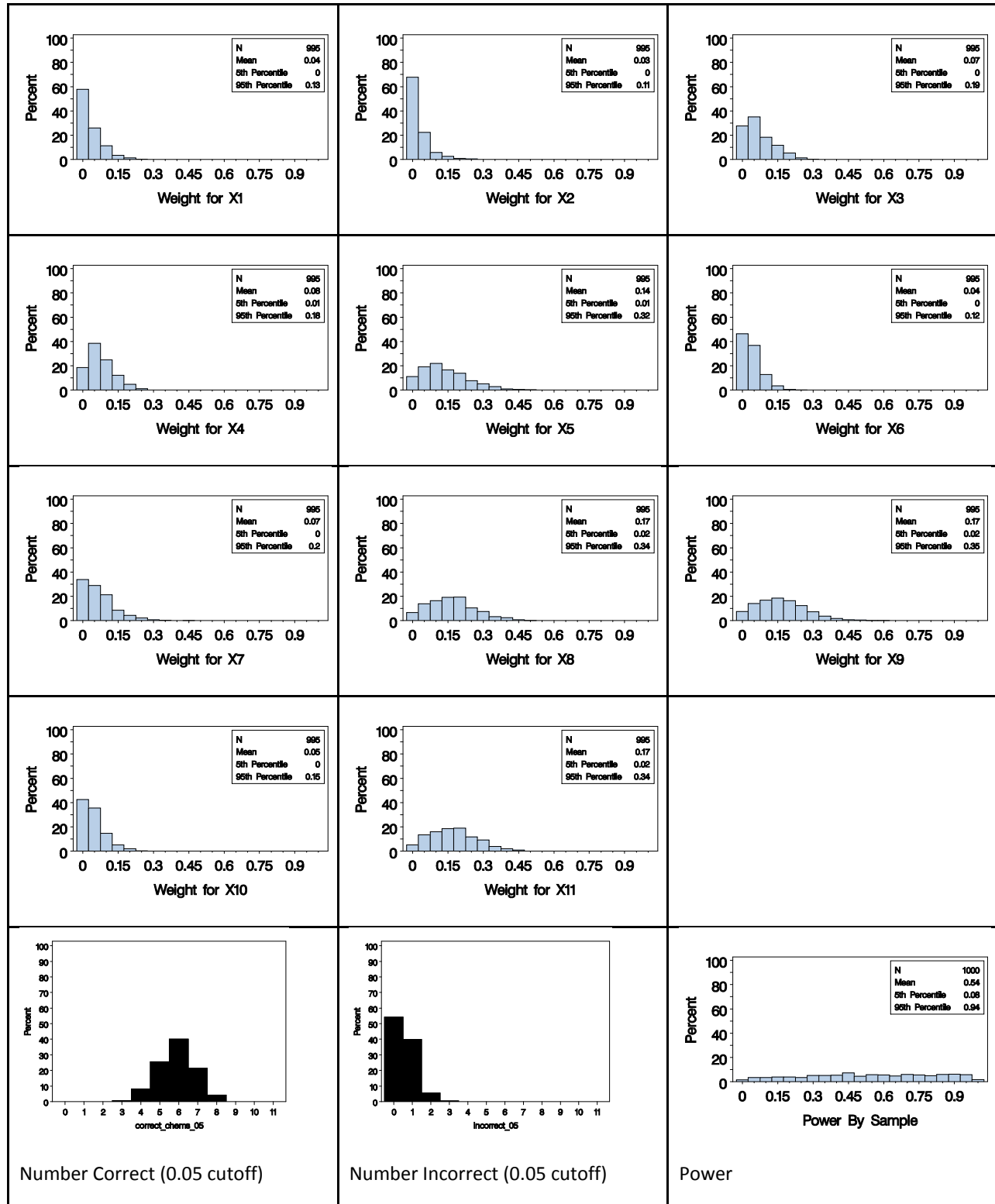


Figure 2.11: Distributions for Weights and Distribution of Power across: 1000 Simulated Datasets; 100 Bootstraps; Sample Size 250 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.3 Correlation with Y

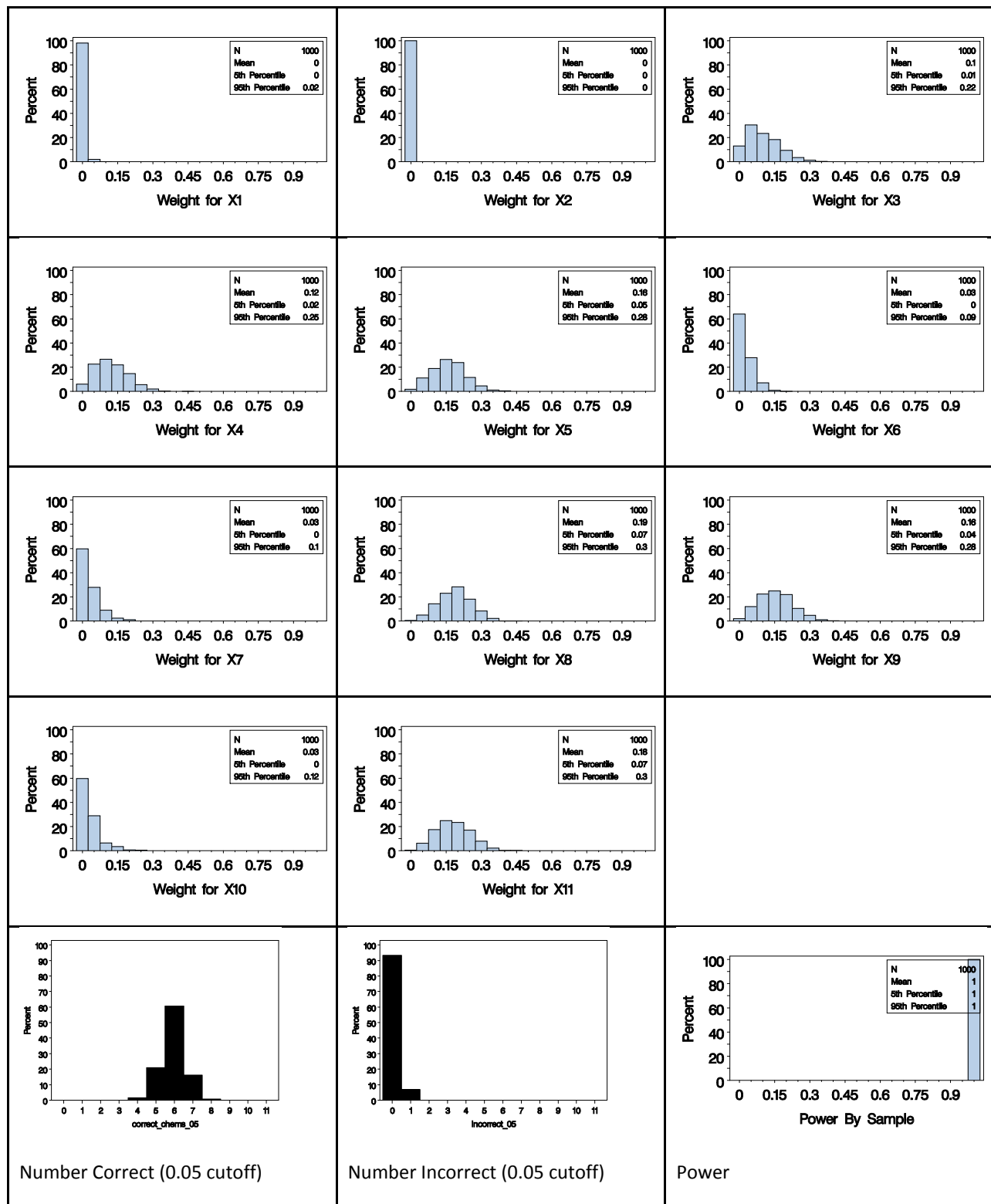
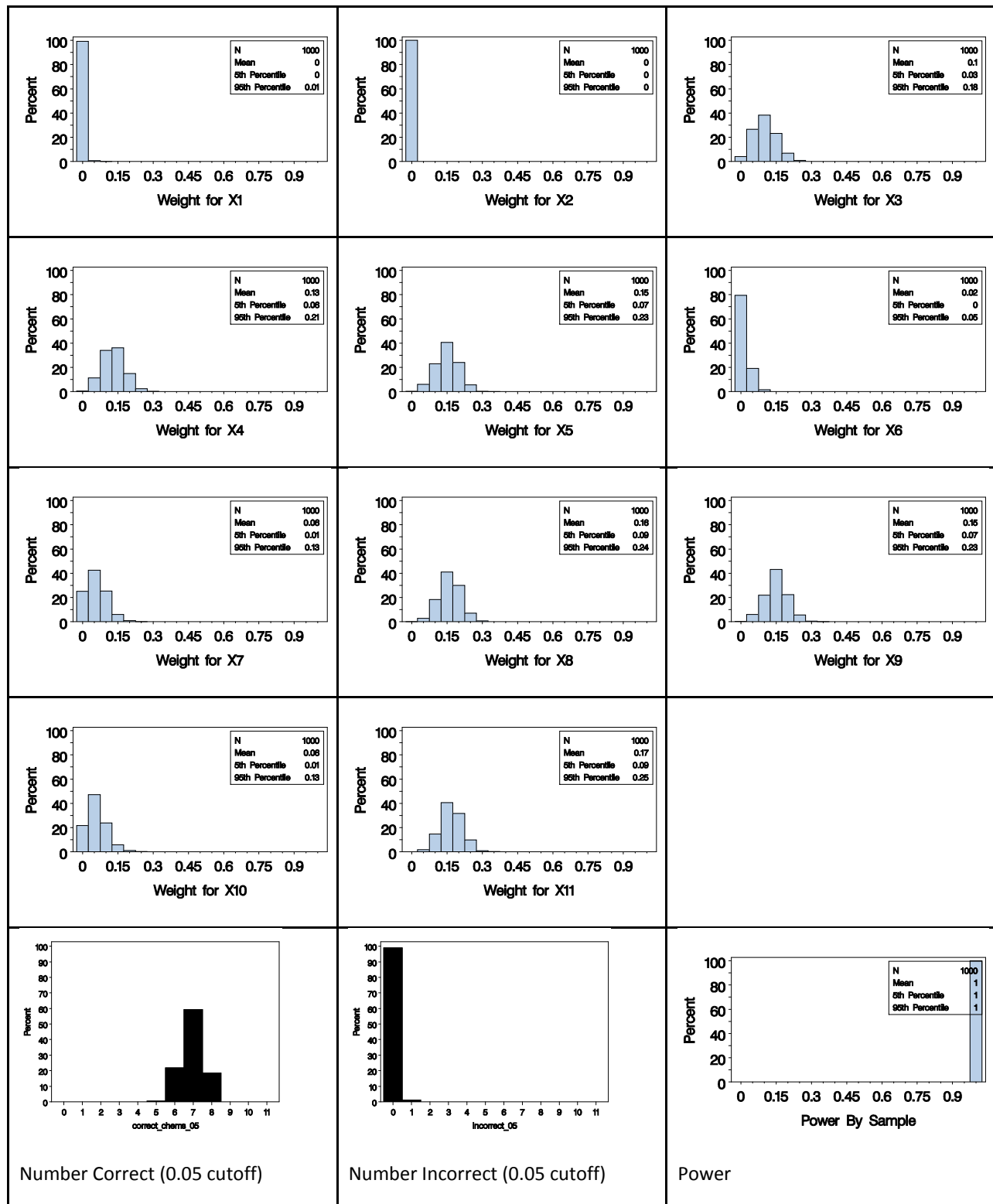


Figure 2.12: Distributions for Weights and Distribution of Power across: 1000 Simulated Datasets; 100 Bootstraps; Sample Size 250 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.3 Correlation with Y; Pairwise Correlations Decreased by 43% (i.e. ridge 1.5)

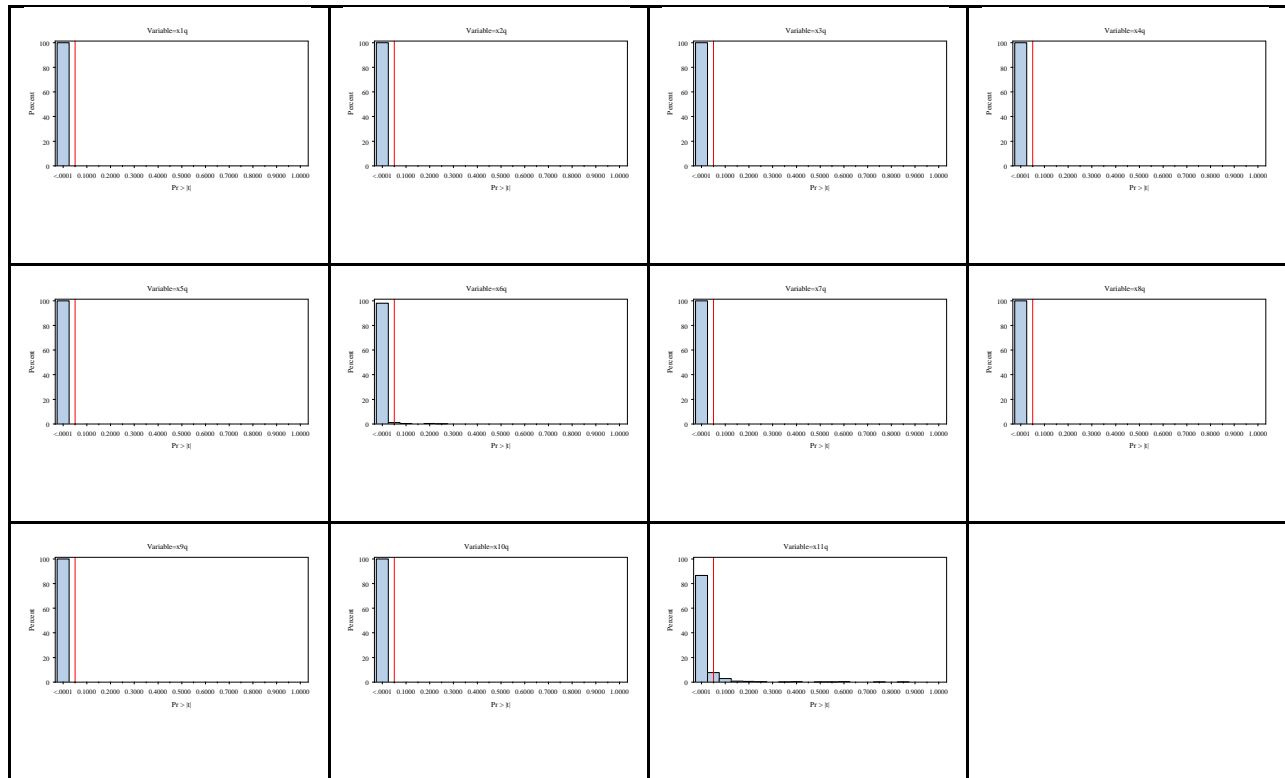


3.5: Traditional Methods Comparison

3.5.1 Ordinary Regression and LASSO Simulations

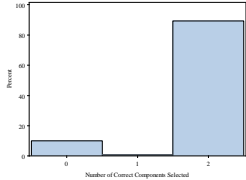
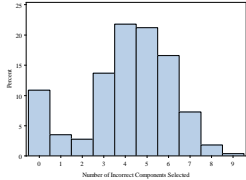
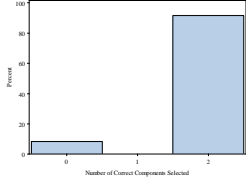
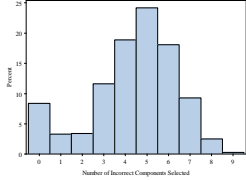
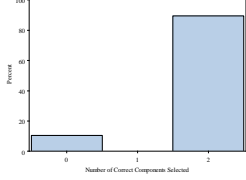
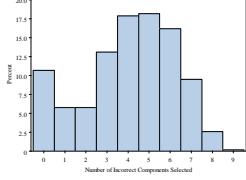
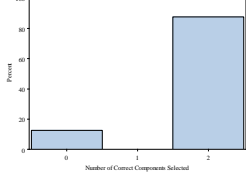
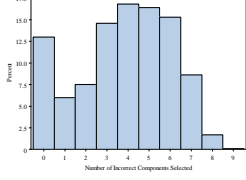
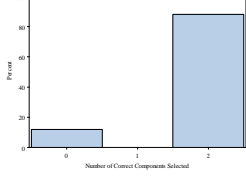
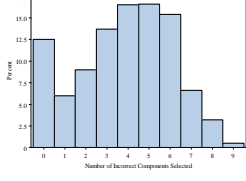
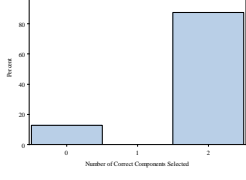
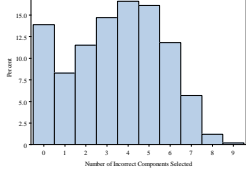
As a comparison for the weighted quartile score, we also performed an ordinary regression analysis (OR) on simulated data. In the OR simulation, we used the simulated data from the case where only one phthalate was correlated with the outcome and that the assumed correlation was 0.3. Figure 2.13 contains the distribution of the p-values for each of the eleven phthalates using the complete data (i.e. not splitting the data like in the weighted analysis case). In these histograms, the red line indicated the 0.05 significance cutoff. That implies that the number of cases to the left of the red line represent the percent of time that a given phthalate is found to be significant in the model. Each of the eleven phthalates were found to be significant in at least 90% of the simulated cases. As anticipated, due to the high correlations, ordinary regression is affected by the multicollinearity. The analysis shows little ability to distinguish between components correlated with the outcome and those that are not, as a result of the complex correlation structure.

Figure 2.13: Ordinary Regression Simulation



We also looked at LASSO as a method for comparison. We used Proc GLMSELECT which employs the LARS algorithm and considers all possible shrinkage parameters and selects the best case. We simulated five cases, all with MHH and MOH correlated with Y at a level of 0.3, but varied the ridge (i.e. reduced the pairwise correlations) from 0 to 0.5 (i.e from 57% of the original correlations to 100% of the original). We found that across all simulation cases, LASSO detected both MHH and MOH correctly in about 80% of simulated cases. However, it also selected several other components in each case. In the case with the highest pairwise correlation between MHH and MOH, it only selected one about 5% of the cases. As the pairwise correlations decreased (i.e. the ridge increased), the distribution of the number of incorrect components shift to the left slightly, but it still tended to select five extra components.

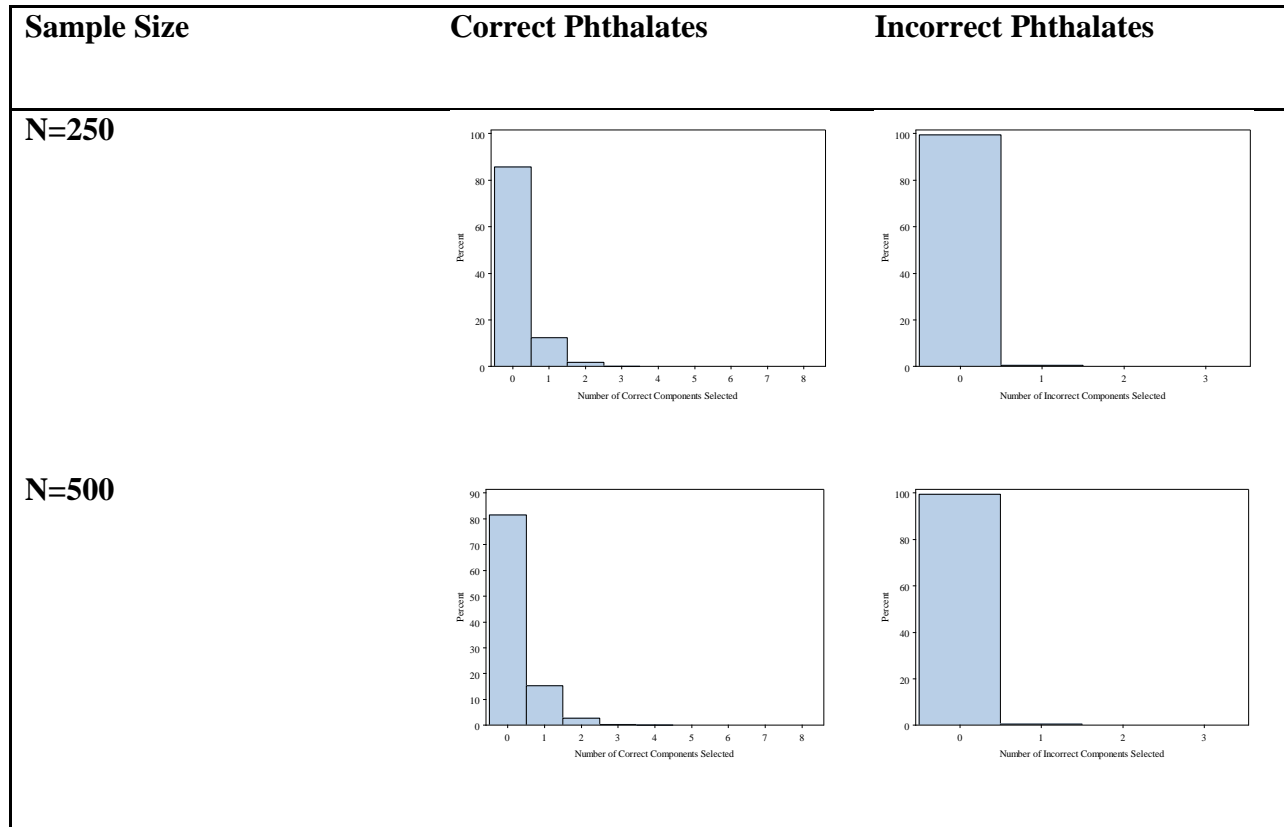
Figure 2.14 LASSO Simulation Results- LARS Algorithm

Ridge (multiplier)	Correlation with Y (Input correlation)	Number of Correct	Number of Incorrect
1 (1)	0.3 (0.3)		
1.1 (0.80)	0.3 (0.375)		
1.2 (0.72)	0.3 (0.417)		
1.3(0.67)	0.3 (0.448)		
1.4 (0.61)	0.3 (0.492)		
1.5 (0.57)	0.3 (0.526)		

3.5.2 Direct Comparison to LASSO

We saw in 3.4 that the weighted quartile score approach has breakdown cases (i.e. where the correlation with the outcome is low and the pairwise correlations are high). In a head-to-head comparison for Corner 1 (where WQS approach demonstrated lowest power, validity and reliability), we simulated 1000 datasets with a sample size of 500, with eight components correlated with the outcome (correlation=0.1) and the phthalate correlation structure. We performed the LASSO analyses with both the split dataset (i.e. 250 observations) and with the complete dataset (i.e. 500 observations). The number of phthalates that should have been assigned weight was eight and a single weighted quartile score analysis with a sample size of 250 was able to detect five. Figure 2.15 presents the histograms of the number of correct and incorrect phthalates assigned weight for both a sample size of 250 and 500.

Figure 2.15: Number of Components Detected by LASSO Correctly and Incorrectly



These results indicate that the LASSO technique probably requires additional data and/or a higher correlation with the outcome to perform since in a majority of the cases, no components were detected. It is clear that this corner is also a breakdown for LASSO, but to a much greater extent than the WQS.

When using LASSO, additional criterion can be used along with the shrinkage. For example, Mallows's C_p criterion can be added to the optimization to balance out under and overfitting. To determine if this additional criterion, or any other criterion available, could offer improvements to the LASSO method, we performed further simulations in Chapter 3.

Conclusions

In a real data case, we may be presented with a small sample size and higher pairwise correlations among the predictors. Through simulations, we have seen the following for WQS approach:

- The weighted quantile score approach has increased stability over ordinary regression
- A single analysis demonstrates lower validity; by adding a bootstrap analysis, validity is improved (See Section 3.4 in Chapter 3)
- The weighted quantile score approach on average does not place weight on components with no correlation with the outcome (high validity)
- Components with high pairwise correlations are assigned relatively lower weights
- There is higher reliability in weights that have lower pairwise correlations
- Increasing sample size is associated with a higher validity and reliability for WQS

From the limited LASSO simulations, we have seen the following:

- LASSO had low validity, as it tended to indicate many components that should not have been selected (Figure 2.14).
- LASSO had lower power in Corner 1 (Figure 2.15)

Overall, the weighted quantile score method is good for a risk analysis setting because it maintains validity and reliability, with improvements as the correlation with the outcome increases. Even in the breakdown case, the bootstrap analysis will indicate on average the “bad actors” more appropriately than other methods at hand (LASSO or regression). When interpreting the weights, one should keep in mind both the pairwise correlations among the components and the correlation with the outcome variable. If the pairwise correlations are high relative to the correlation with the outcome, there may be a “breakdown case.” In that setting, the weights should be considered in conjunction with the pairwise correlations a given component has with other components. If a component has a minimal weight (i.e. less than 0.05 or 0.01 if a large number of components or complex correlation structure) and is highly correlated with other components assigned minimal weight, the two are likely important, but have smaller weights as a

result of their high pairwise correlation. From this type of analysis, we are able to detect components that are associated with a given health outcome and assess the total body burden they impose on an individual. Figure 2.6 demonstrates the areas where the WQS performs best (green area) and an analyst should be aware of his or her placement on this spectrum.

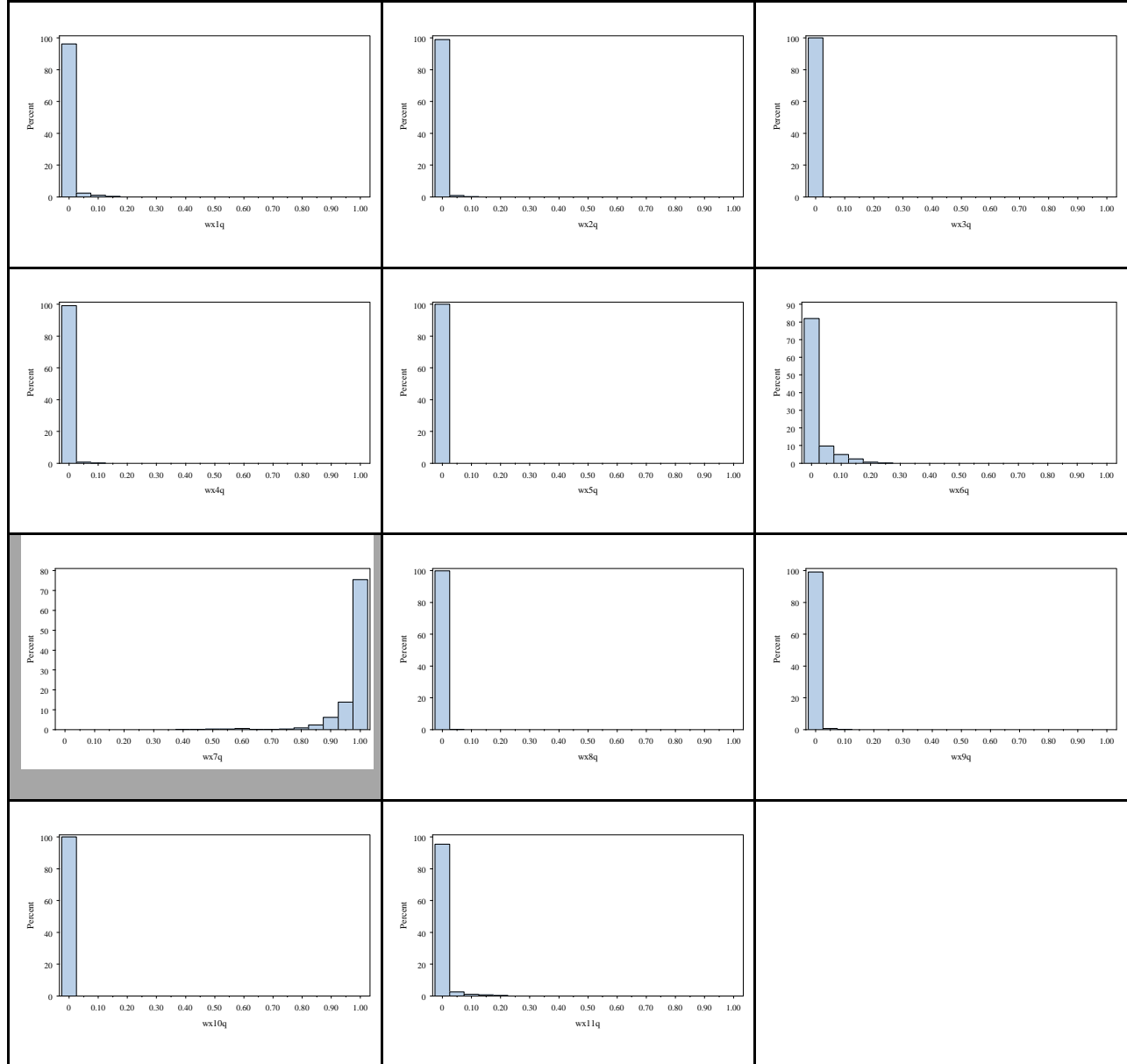
III. Appendix: Supplementary Material to Chapter 2

Introduction

This chapter provides supplementary material to Chapter Two, primarily the results of further simulations with different correlation structures and sample sizes. These simulations further demonstrate the validity and reliability of the weighted quantile score under different conditions. Each figure contains the simulation conditions and a brief summary and interpretation of the results. Overall, the simulations show that in ideal settings (high correlation with outcome, lower pairwise correlations, and large sample size), the method performs with high validity and reliability. They also show that in settings with more complex correlation structures (large number of components correlated with outcome, high pairwise correlations, low correlation with outcome, etc) that the method does have lower validity and reliability, notably measured through an increased false negative rate. We demonstrated in Chapter 2 that a bootstrap analysis lessens these effects, and that the method still outperforms LASSO and regression in the same cases. Also included is a demonstration of LASSO with other criterion. The final component of this chapter is a demonstration of the weight quantile score and the improvement that the addition of a bootstrap provides.

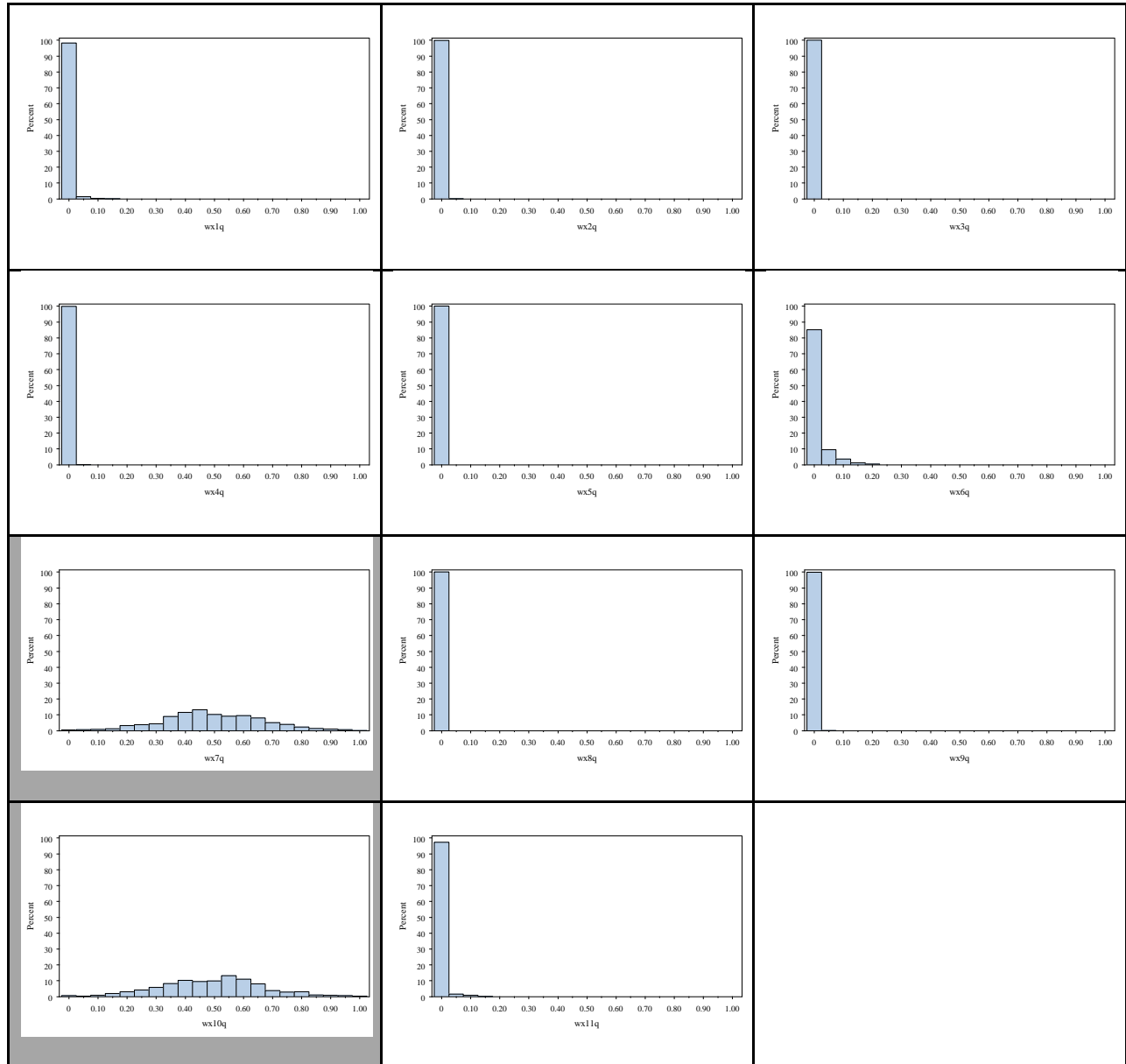
3.1 Various Simulations for Weighted Quantile Score with Phthalate Correlation Structure from Figure 1.1 in Chapter 2

3.1: MHH (X7) Correlated with Y (Corr=0.3), Sample Size 1000, Observed Phthalate Correlation Structure (Ch 2 Figure 1.1), No Other components Correlated with Outcome



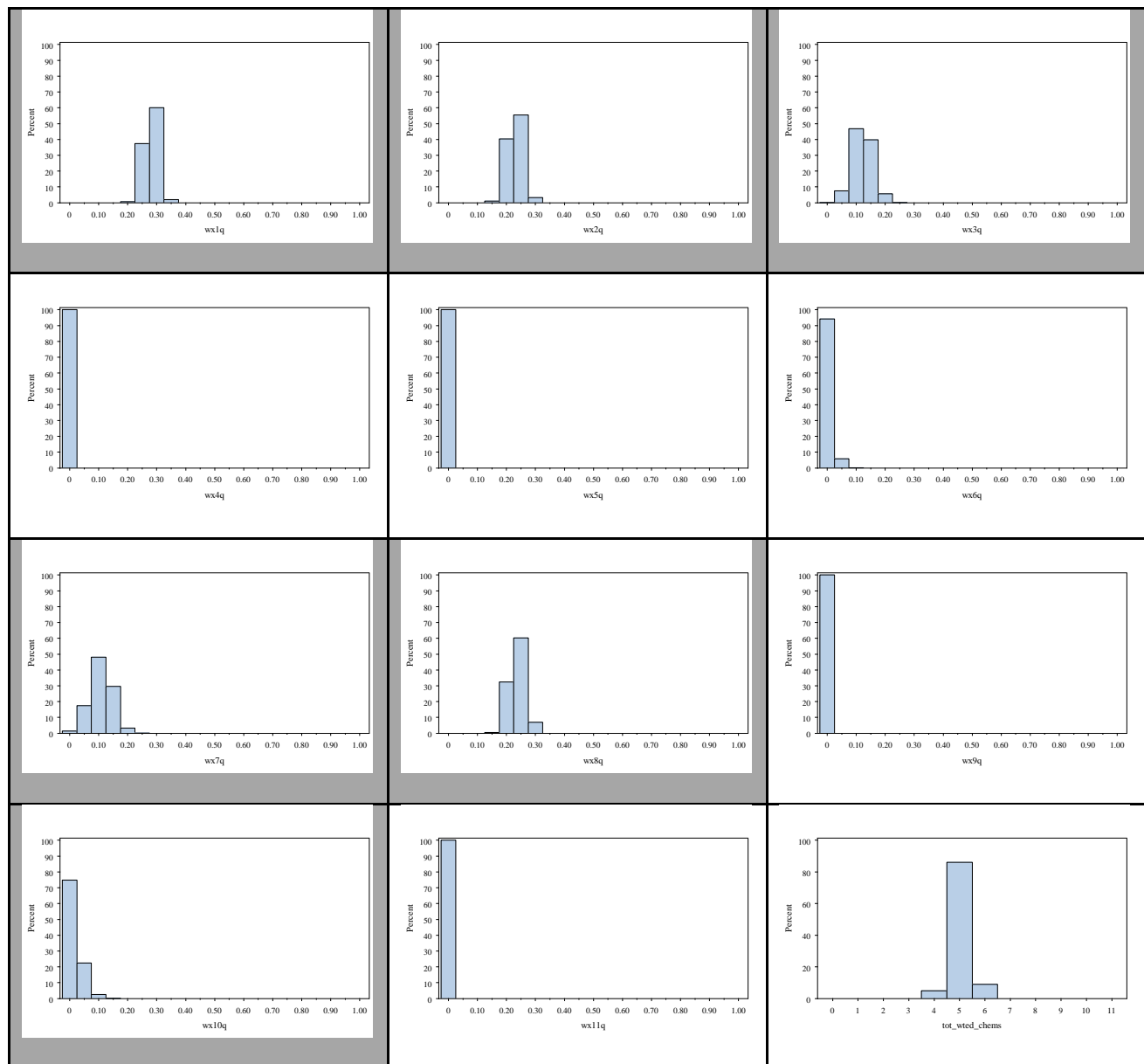
This simulation was done to determine if given a high pairwise correlation between two components (X7 and X10 have correlation of 0.92), if only one is correlated with the outcome, could the WQS distinguish between the two components? Because it is clear that the weight is solely placed on X7, the method was able to distinguish which component is correlated with the outcome and which was not. It also does not place more than marginal weight on any other components.

3.2 MHH (X7), MOH (X10) Correlated with Y (Corr=0.3), Pairwise Correlation=0.92, Sample Size 1000, Horizontal Axis is 0 to 1 and Vertical Axis 0 to 100 for All Histograms



This simulation was done to determine if given a high pairwise correlation between two components (X7 and X10 have correlation of 0.92), if both are correlated with the outcome, could the WQS detect both components or just one. Because it is clear that the weight is placed evenly on X7 and X10, the method was able to detect that both are related to the outcome. It also does not place more than marginal weight on any other components.

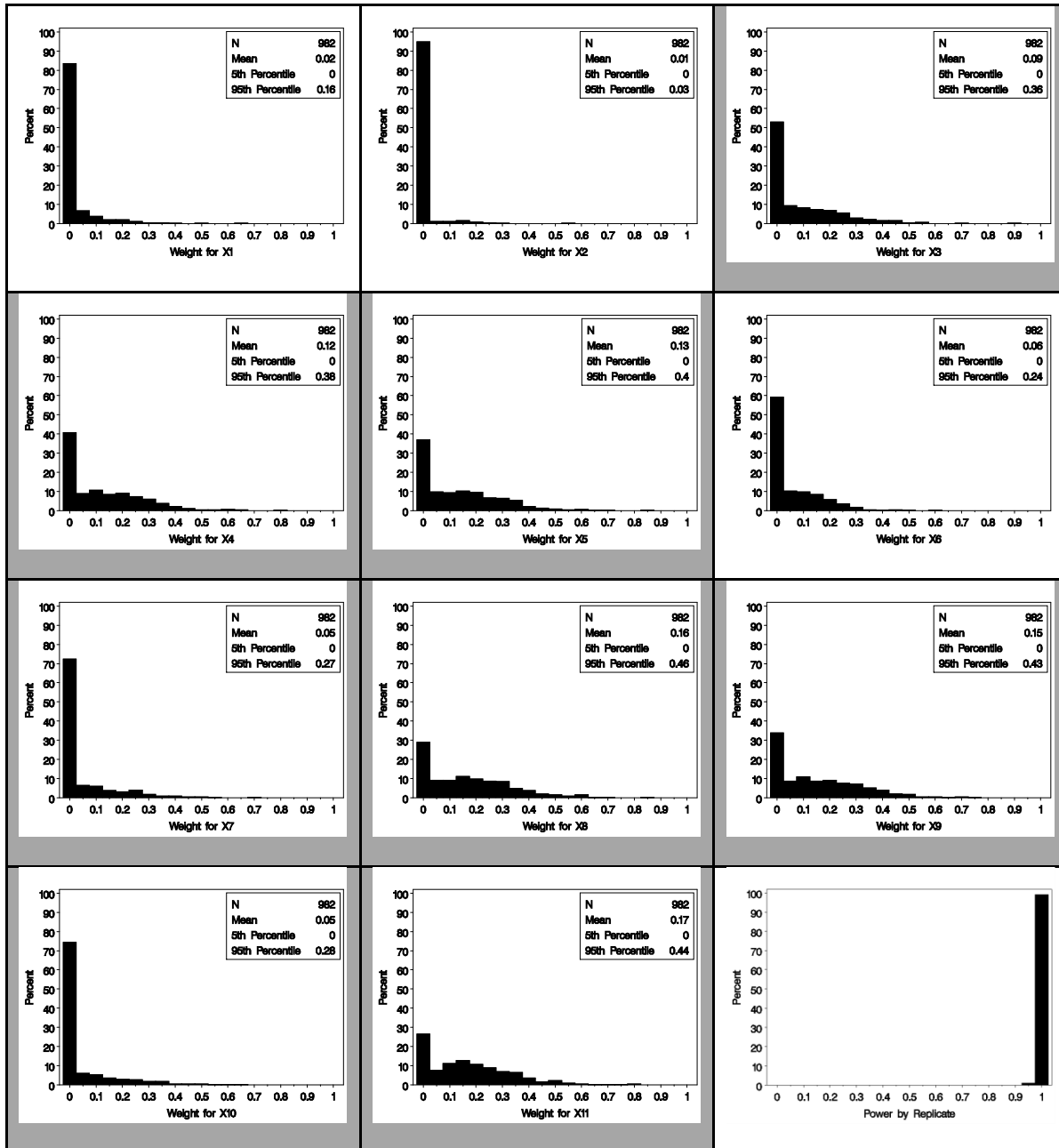
3.3 All Six High Molecular Weight Phthalates Correlated with Y of 0.5; Observed Phthalate Correlation (Figure 1.1 in Ch 2); Note: Highest pairwise correlations occur between high molecular weight phthalates; Final histogram is the distribution of the number of weights greater than 0.05; Components correlated with outcome were: X1, X2, X3, X7, X8, X10 and are indicated in the table with shaded background.



This simulation shows that the method occasionally missed one of the size components despite the high correlation with the outcome. However, a “miss” was defined as a weight less than 0.05 not a true 0 and the bootstrap analysis was not performed in this case. Additionally, the distribution of X10, the component that was frequently “missed” still had a distribution that was different from the remaining five components. The components that had no correlation(X4, X5, X6, X9, X11) with the outcome in the simulation had near perfect distribution of 0 weight. The pairwise correlations between X10 and X3 and X7 are 0.85 and 0.92. As shown in Ch 2, components with high pairwise correlations can have lower estimated weights. So the near zero weights for X10 do not indicate a lack of importance, but high multicollinearity.

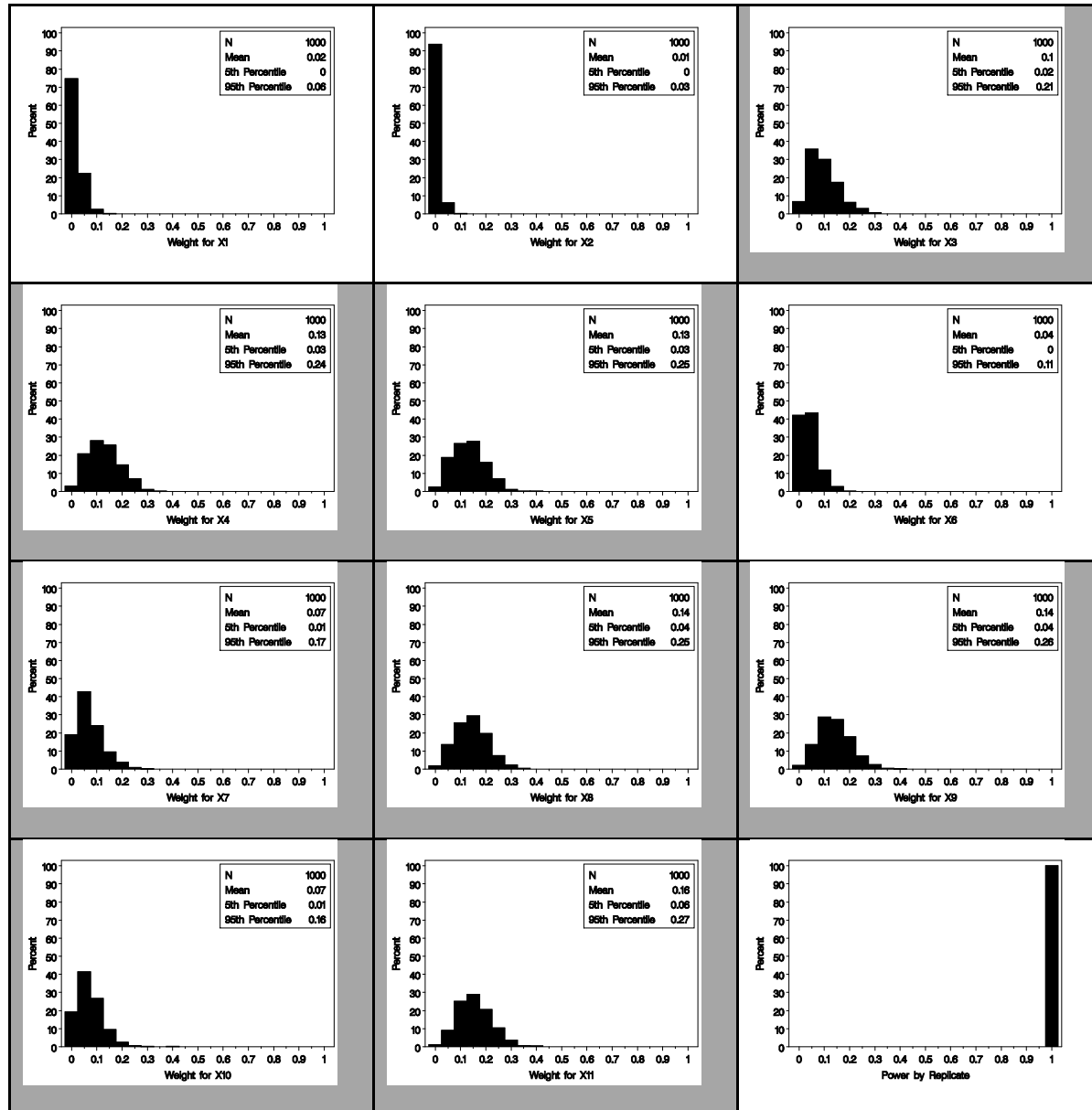
3.2 Simulated Bootstrap Analyses For Breakdown Cases with Increased Sample Size

3.4 Distributions for Weights and Distribution of Power across: 1000 Simulated Datasets; 100 Bootstraps; Sample Size 500 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.1 Correlation with Y; No Ridge- Pairwise Correlations from Figure 1.1 in Ch 2.



These results show improvement over Figure 4.5 (with only change the increased sample size). The weights for X1, X2 and X6, which should be zero, decrease for X1 and X2 by about half (0.04 to 0.02). The average weight for X6 actually increased, but the distribution tended much closer to 0- therefore having a lower false positive rate. The weights for variables that were assigned smaller weight than expected (X7, X10) also increased from 0.02 to 0.05 and 0.03 to 0.05 respectively. While this is still a breakdown case due to the high pairwise correlations and low correlations with the outcome, the increased sample size did marginally improve the validity and reliability.

3.5. Distributions for Weights and Distribution of Power across: 1000 Simulated Datasets; 100 Bootstraps; Sample Size 500 for Weight Estimation; X1, X2, X6 NOT Correlated with Y; Remaining components 0.1 Correlation with Y; Ridge 1.5- Pairwise Correlations from Figure 1.1 Reduced by 43%



These results show improvement over Figure 3.4 (with only change the ridge of 1.5). The average weights for X1 and X2 are the same, but the range has decreased (i.e. reliability is higher). For the simulation in Figure 3.4, the 95th percentile for X1 was more than twice what it was in Figure 3.4. For X6, which also should have been estimated 0, the estimate was 0.06 with a 95th percentile of 0.24, versus 0.04 and 0.11 respectively in this simulation case. This simulation case also has a much lower false negative rate than Figure 3.4. For many of the components in Figure 3.4, the false negative rate is at least 20%, but in this case it is below 10% for most and very near zero for many.

3.3 Comparison to LASSO

3.6 LASSO Results: No Selection Criterion; LARS Algorithm

Simulation Setting	Number of Correct	Number of Incorrect
<p>X3, X7 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 2</p>		
<p>X3, X7, X10 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 3</p>		
<p>All Six High Molecular Weight Phthalates Correlated with Y of 0.5; Observed Phthalate Correlation (Figure 1.1 in Ch 2); Note: Highest pairwise correlations occur between high molecular weight phthalates; Components correlated with outcome were: X1, X2, X3, X7, X8, X10</p> <p>Number Correct Should be: 6</p>		

3.7 ADJRSQ Adjusted R-square statistic

Simulation Setting	Number of Correct	Number of Incorrect
<p>X3, X7 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 2</p>		
<p>X3, X7, X10 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 3</p>		
<p>All Six High Molecular Weight Phthalates Correlated with Y of 0.5; Observed Phthalate Correlation (Figure 1.1 in Ch 2); Note: Highest pairwise correlations occur between high molecular weight phthalates; Components correlated with outcome were: X1, X2, X3, X7, X8, X10</p> <p>Number Correct Should be: 6</p>		

3.8 AIC Akaike information criterion

Simulation Setting	Number of Correct	Number of Incorrect
<p>X3, X7 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 2</p>		
<p>X3, X7, X10 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 3</p>		
<p>All Six High Molecular Weight Phthalates Correlated with Y of 0.5; Observed Phthalate Correlation (Figure 1.1 in Ch 2); Note: Highest pairwise correlations occur between high molecular weight phthalates; Components correlated with outcome were: X1, X2, X3, X7, X8, X10</p> <p>Number Correct Should be: 6</p>		

3.9 AICC Corrected Akaike information criterion

Simulation Setting	Number of Correct	Number of Incorrect
<p>X3, X7 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 2</p>		
<p>X3, X7, X10 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 3</p>		
<p>All Six High Molecular Weight Phthalates Correlated with Y of 0.5; Observed Phthalate Correlation (Figure 1.1 in Ch 2); Note: Highest pairwise correlations occur between high molecular weight phthalates; Components correlated with outcome were: X1, X2, X3, X7, X8, X10</p> <p>Number Correct Should be: 6</p>		

3.10 CV Predicted residual sum of square with k -fold cross validation

Simulation Setting	Number of Correct	Number of Incorrect
<p>X3, X7 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 2</p>		
<p>X3, X7, X10 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 3</p>		
<p>All Six High Molecular Weight Phthalates Correlated with Y of 0.5; Observed Phthalate Correlation (Figure 1.1 in Ch 2); Note: Highest pairwise correlations occur between high molecular weight phthalates; Components correlated with outcome were: X1, X2, X3, X7, X8, X10</p> <p>Number Correct Should be: 6</p>		

3.11 PRESS Predicted residual sum of squares

Simulation Setting	Number of Correct	Number of Incorrect
<p>X3, X7 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 2</p>		
<p>X3, X7, X10 Correlated with Y, Correlation=0.3; All remaining components not correlated with outcome; Observed correlation from Figure 1.1 Ch 2.</p> <p>Number Correct Should be: 3</p>		
<p>All Six High Molecular Weight Phthalates Correlated with Y of 0.5; Observed Phthalate Correlation (Figure 1.1 in Ch 2); Note: Highest pairwise correlations occur between high molecular weight phthalates; Components correlated with outcome were: X1, X2, X3, X7, X8, X10</p> <p>Number Correct Should be: 6</p>		

The LASSO simulations all essentially demonstrate a similar conclusion: while the method is able to detect the “right” components a majority of the time, it tends to pick up additional components across the simulation cases. There are criterion that tend to perform better than others, but in the cases shown, it seems as though the LARS algorithm is likely the best option since the distribution of incorrect components is more normally distributed than it is left skewed, which is how the other criteria tend to perform.

3.4 Applying to Real Data: Demonstration of Development of WQS

3.4.1 Problems in Real data

As shown in Chapter 2, the weighted quartile score performs best in cases with higher correlations with the outcome relative to the pairwise correlations of the components. But in a real data case, we might be presented with limited sample size and pairwise correlations among the predictors that are larger than the correlations with the outcome. In that case, there is reason to have concern about the performance of the weighted quartile score. In this breakdown case, we propose the use of bootstrapping to improve the results from a single analysis.

3.4.2 Phthalate Breakdown Case Demonstration

We demonstrate the application of the bootstrap analysis to the case with a total sample size of 500 (i.e. 250 for the estimation of the weights and 250 for the validation step). We assume again that there are eight components correlated with the outcome and that correlation is set to 0.1 and the pairwise correlations are as observed in the NHANES dataset which are listed in Figure 1.1. In the simulated sample, the weighted quartile score approach assigned nonzero weight to five of the pre-specified eight components. These weights are given in Table 3.1, where the components that should have been assigned weight are indicated with an asterisk.

Table 3.1: Results from Breakdown Case with Phthalates with Sample Size of 250 and 8

Phthalates Marked with Asterisk Correlated with Y at 0.1 Level

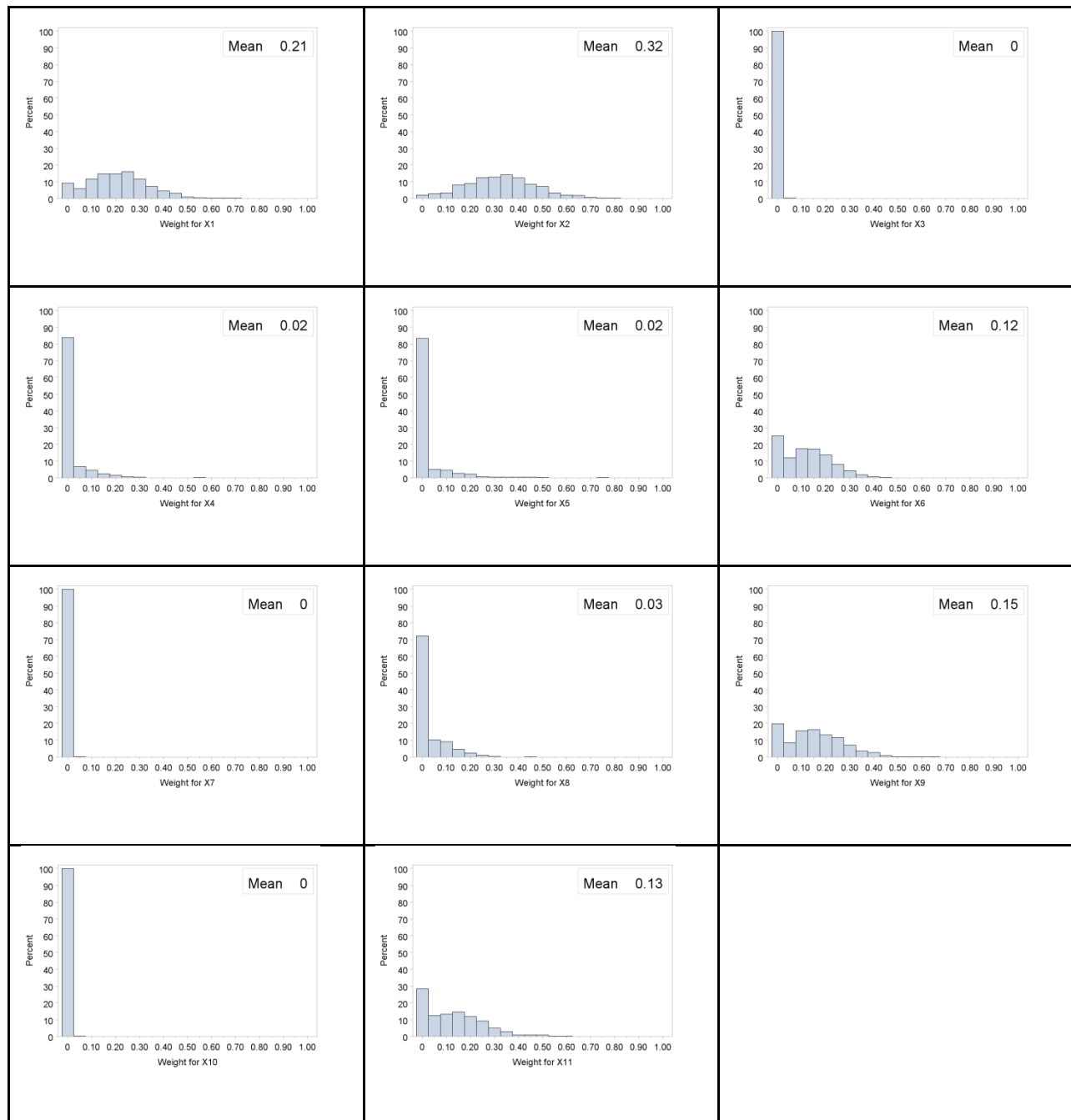
Phthalate	Weight
CNP	0.19*
COP	0.09*
ECP	0.00
MBP	0.00*
MC1	0.00*
MEP	0.16*
MHH	0.00
MHP	0.38*
MIB	0.00*
MOH	0.00
MZP	0.17*

The index created from these weights was validated in the second half of the dataset and found to be significant (estimate: 4.58; p-value= 0.002). So while the model is significant and the approach detected five of the eight phthalates, we would like for it to be able to detect all eight phthalates and to be validated as significant in the validation dataset.

The results show that the method was sensitive (i.e. it did not assign weights to components that are not associated with the outcome) but that it was not as specific (i.e. it missed components that should have been indicated). We propose that by taking bootstrap simulations from the original data, we will be able to improve specificity and detect the components that are

missed in a single analysis. We took 1000 random bootstrap samples from the single simulated dataset and estimated the weights using the Trust Region algorithm for each sample. The distributions of the weights are given in Figure 3.12; the average weight is given in the inset of each histogram.

Figure 3.12: Distribution of Weights from 1000 Bootstrap Samples of size 250 from Data from Table 3.1



When the average bootstrap sample weight is calculated, there is nonzero value placed on all eight of the pre-specified components. We also applied this average weighted quartile score to the validation dataset and found that it was significant (estimate: 6.80; p-value<0.0001).

IV. Application of Method: Environmental Chemicals, Non-Chemical Stressors and Liver Health

1. Introduction

As the rate of obesity continues to reach staggering levels, diseases that are associated with obesity are also on the rise. One such disease is non-alcoholic fatty liver disease (NAFLD) which is also referred to as non-alcoholic steatohepatitis (NASH) in more severe cases. NAFLD and NASH are characterized by fatty deposits in the liver along with inflammation and damage (NDDIC). NAFLD is the most common liver disease in the United States and the trend is followed worldwide, to an extent that could be deemed epidemic (Cave, 2007). It is believed that NASH/NAFLD are associated with increased visceral adiposity (large waist circumference) plus insulin resistance (Krawczyk, et al 2010). These conditions are both known to be caused by obesity and poor nutrition. In addition, oxidative stress has been implicated as a cause of NASH/NAFLD (Krawczyk, et al 2010). The most common biomarker for NAFLD is elevated transamines, especially alanine amino-transferase (ALT) (Cave, 2007). Environmental chemicals have been linked to an increase in ALT and/or NASH/NAFLD including occupational exposure to petrochemicals (Cave, 2007) and polychlorinated biphenyls (PCBs), dioxins, furans and heavy metals (Christensen, 2013).

In risk assessment focus was shifted to such “non-chemical stressors” after the National Research Council’s report in 2009 recommended the consideration of both chemical and non-chemical stressors on public health (Lewis, 2011). Common non-chemical stressors include low socio-economic status, race/ethnicity, obesity, and occupational and community-related exposures. These non-chemical stressors are also referred to as “vulnerability factors” (Lewis, 2011). Because NASH/NAFLD, and fatty liver disease in general, are highly influenced by

obesity, poor nutritional status could likely be a non-chemical stressor for NASH/NAFLD. In this paper, we will further investigate the effect of the environmental chemicals based on the methods from Carrico, et al (2013) and investigate a nutritional index for non-chemical stressors. A positive relationship (i.e. increase in ALT) between ALT and the nutrition index we consider representative of non-chemical nutritional stressors. If the relationship is negative (i.e. decrease in ALT), then the index places weight on the nutritional components that are non-stressors on liver functioning.

2. Methods

2.1 Description of Data

The NHANES studies are a series of studies conducted by the CDC to assess the health and nutritional status of a representative sample from the US population, including both adults and children (CDC). The data are publically available on the CDC website and contain data collected from a personal interview and a physical examination which includes the collection of biological specimens (blood and urine). The 2003-2004 cycle is the most recent cycle that contains blood serum data on polychlorinated biphenyls (PCBs), dioxins, furans and heavy metals. Data from both the interview and the physical examination were used in this analysis. From the interview, data from the two 24-hour total dietary recalls along with demographic data (age, gender, income to poverty ratio, body mass index (BMI), and race/ethnicity) were used. Data on blood serum levels of coplanar PCBs, noncoplanar PCBs, dioxins and furans, and heavy metals were considered. In total, 34 PCBs, dioxins and furans; 3 heavy metals; and 56 nutrients were used in the analysis and are listed in Table 4.1a-b. Each nutrient was summed across the two days, adjusted for total caloric intake, and then scored into quartiles. Lipid-adjusted blood serum analyte levels for PCBs, dioxins, furans and heavy metals were also scored into quartiles.

The data used are a subsample of 928 subjects from the 2003-2004 NHANES dataset, after subjects with missing data for the environmental chemicals or the dietary recall are excluded. In addition, following Christensen, et al (2013), subjects with history or indication of liver disease, indication of Hepatitis B, Hepatitis C, or high alcohol intake are excluded from the analysis.

Table 4.1a: Analytes Considered in Analyses

ANALYTES			
Dioxin-like compounds	Non-dioxin-like PCBs		Metals
PCB 28	PCB 44	PCB 153	Cadmium
PCB 66	PCB 49	PCB 170	Lead
PCB 74	PCB 52	PCB 177	Mercury
PCB 105	PCB 87	PCB 178	
PCB 118	PCB 99	PCB 180	
PCB 156	PCB 101	PCB 183	
1,2,3, 6,7,8-HXCDD	PCB 110	PCB 187	
1,2,3,4,6,7,8-HPCDD	PCB 138	PCB 194	
1,2,3,4,6,7,8,9-OCDD	PCB 146	PCB 196	
1,2,3,4,6,7,8-HPCDF	PCB 149	PCB 199	
3,3',4,4',5-PNCB	PCB 151	PCB 206	
		PCB 209	

Table 4.1b: Dietary Nutrients Considered in Analyses

Nutrients		
Vitamins/Minerals	Fats	Others
Vitamin E	Total Fat	Protein
Vitamin A	Total Saturated Fatty Acids (SFA)	Carbohydrates
Alpha Carotene	SFA 4:0 (Butanoic)	Total Sugars
Beta Carotene	SFA 6:0 (Hexanoic)	Dietary Fiber
Beta Cryptoxanthin	SFA 8:0 (Octanoic)	Cholesterol
Lycopene	SFA 10:0 (Decanoic)	Sodium
Lutein+zeaxanthin	SFA 12:0 (Dodecanoic)	Caffeine
Thiamin	SFA 14:0 (Tetradecanoic)	Theobromine
Riboflavin	SFA 16:0 (Hexadecanoic)	
Niacin	SFA 18:0 (Octadecanoic)	
Vitamin B6	Total Monounsaturated Fatty Acids (MFA)	
Folate	MFA 16:1 (Hexadecenoic)	
Folic Acid	MFA 18:1 (Octadecenoic)	
Food Folate	MFA 20:1 (Eicosenoic)	
Vitamin B12	MFA 22:1 (Docosenoic)	
Vitamin C	Total Polyunsaturated Fatty Acids (PFA)	
Vitamin K	PFA 18:2 (Octadecadienoic)	
Calcium	PFA 18:3 (Octadecatrienoic)	
Phosphorus	PFA 18:4 (Octadecatetraenoic)	
Magnesium	PFA 20:4 (Eicosatetraenoic)	
Iron	PFA 20:5 (Eicosapentaenoic)	
Zinc	PFA 22:5 (Docosapentaenoic)	
Copper	PFA 22:6 (Docosahexaenoic)	
Potassium		
Selenium		

The outcome variable of interest was serum alanine amino-transferase (ALT) level, which is an indication of overall liver health, and a preliminary test for fatty liver disease. A high level of ALT is indicative of poor liver health. The distribution of ALT in the dataset was right skewed and therefore the outcome variable was modeled as the natural log of ALT ($\text{Log}(\text{ALT})$), which was normally distributed.

Along with controlling for the analytes and nutrients, demographic variables were also considered in the analyses. These included gender (male, female), age at time of interview (in years), race/ethnicity (dichotomized to Non-Hispanic White and Others), poverty status (ratio of family income to poverty threshold), and BMI. For poverty status, any ratio greater than or equal to five was given the maximum value of 5.

For the analyses, the complete dataset was randomly divided into two groups: one was used to estimate the weights (referred to as the “test” dataset) and one was used to validate (referred to as the “validation” dataset) these results. Chapter 2 extended the weighted quantile score approach used by Christensen, et al to include a bootstrap analysis. Carrico, et al demonstrated improved accuracy (in terms of validity and reliability) by defining the weighted score as that formed by the average bootstrap weights. Therefore, the test dataset was used to generate 1000 bootstrap samples, from which an “environmental chemical score” (ECS) and a “nutritional stressor score” (NSS) were determined by the average weights. These indices were validated in the single validation dataset. Details are provided in Section 2.2.

2.2 Preliminary Statistical Analyses

All analyses were performed in SAS 9.2. Prior to the beginning of the analysis, some preliminary checks on the data were performed. Pairwise correlations between the PCBs, dioxins, furans, and heavy metals and the dietary nutrients were calculated in order to

demonstrate the complex correlation structure, and the resulting multicollinearity problem. The distribution of the outcome variable, log(ALT), was also checked to insure it was normally distributed.

Prior to estimating the weights, the core model was determined. Covariates of interest included age (in years), gender (male/female), a binary race variable (Non-Hispanic White vs Others), BMI (continuous), and Poverty:Income ratio (continuous; all ratios greater than 5 scored as 5). All continuous variables were checked to determine if higher order terms needed to be included in the model.

2.3 Weighted Quartile Scores

Our objective is the formation of a weighted index for the analytes, ECS, and the dietary nutrients, NSS, that maximizes the likelihood for a multiple regression model predicting the mean(μ) of Log(ALT). Following Carrico, et al (2013), ECS and NSS were calculated using the average weights from the 1000 bootstrap samples where the sample weights were significant. Due to the large number of components in both indices, ECS and NSS were estimated separately. The unknown parameters were estimated in each bootstrap sample:

$$\text{Log}(\mu) = \beta_0 + \beta_1 * \sum_{i=1}^{37} w_i * p_i + \mathbf{z}^* \boldsymbol{\varphi}$$

$$\text{Log}(\mu) = \beta_0 + \beta_2 * \sum_{i=1}^{56} u_i * v_i + \mathbf{z}^* \boldsymbol{\varphi}$$

2.1

Where:

\mathbf{w} is the 37x1 vector of weights, w_i is the weight for the i^{th} chemical, p_i

\mathbf{u} is a 56x1 vector of weights; u_i is the weight for the i^{th} nutrient, v_i

The covaraites of interest (age, gender, race/ethnicity, BMI, and poverty index) are accounted for in the vector \mathbf{z} with corresponding parameter in vector $\boldsymbol{\phi}$. The common parameters are not constrained to be equal for all models (i.e. β_0 is estimated for each model individually and despite notation is not assumed to be equal across all models; the same is true for ϕ , β_1 and β_2).

Using the weights from the bootstrap samples that are associated with a significant β_1 or β_2 in each bootstrap test dataset, ECS and NSS are calculated as:

$$\begin{aligned} \text{ECS} &= \sum_{i=1}^{37} \bar{w}_i * p_i \\ \text{NSS} &= \sum_{i=1}^{56} \bar{u}_i * v_i \end{aligned} \tag{2.2}$$

The models in 2.3 are then estimated and tested in the validation dataset. For clarification, the weights from each bootstrap sample are validated back in the bootstrap sample and the final average bootstrap weights are validated in the single validation dataset.

$$\begin{aligned} \text{a. } \text{Log}(\text{ALT}) &= \beta_0 + \beta_1 * \text{ECS} + \mathbf{z}' * \boldsymbol{\phi} \\ \text{b. } \text{Log}(\text{ALT}) &= \beta_0 + \beta_2 * \text{NSS} + \mathbf{z}' * \boldsymbol{\phi} \\ \text{c. } \text{Log}(\text{ALT}) &= \beta_0 + \beta_1 * \text{ECS} + \beta_2 * \text{NSS} + \mathbf{z}' * \boldsymbol{\phi} \\ \text{d. } \text{Log}(\text{ALT}) &= \beta_0 + \beta_1 * \text{ECS} + \beta_2 * \text{NSS} + \beta_{12} * \text{ECS} * \text{NSS} + \mathbf{z}' * \boldsymbol{\phi} \end{aligned} \tag{2.3}$$

When β_1 in 2.3a and β_2 in 2.3b are significant then model 2.3d is fit. When the interaction term in 2.3d is not significant then the final model, 2.3c, is estimated. If 2.3d has a significant interaction term, 2.3c is not fit. Parameters are not constrained to be equal for all models (i.e. β_0

is estimated for each model individually and despite notation is not assumed to be equal across all models; the same is true for ϕ , β_1 and β_2).

Results

3.1 Preliminary Results

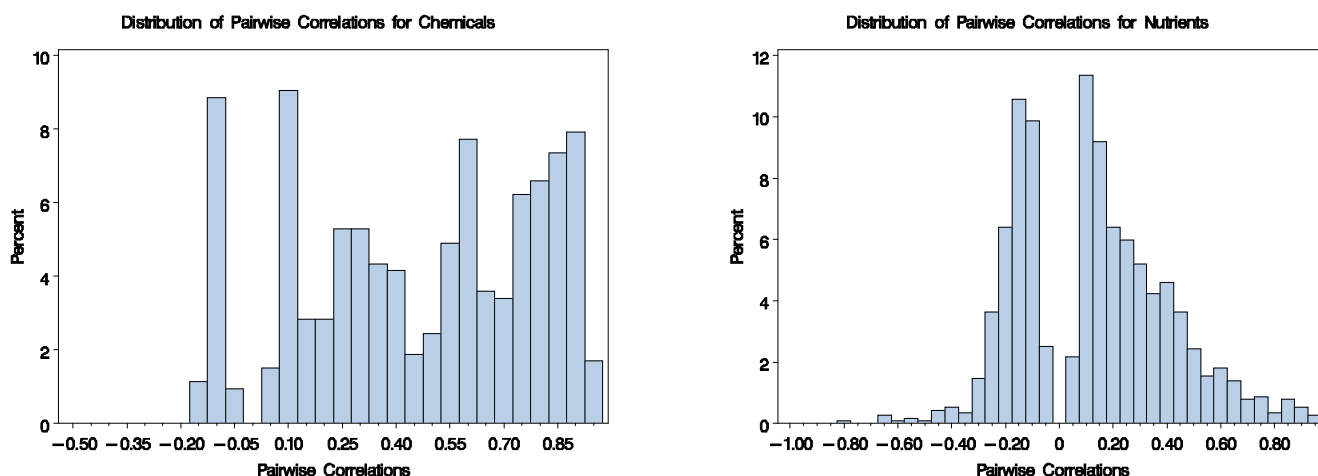
The data were randomly split into two groups in order to have a test and a validation dataset. The test portion of the data had a resulting sample size of 464, with the remaining 464 observations allocated to the validation dataset. The two subsets were compared across the covariates and the outcome. There were no significant differences between the test and the validation dataset.

Table 4.2: Comparison of Test and Validation Datasets for Covariates and Outcome

Variable	Test		Validate		P-value
Log(ALT) (Mean, SD)	2.98	0.41	2.94	0.36	0.67
Age (Mean, SD)	38.03	24.00	35.96	24.15	0.85
Gender (%Female)	55.2%		51.9%		0.32
Race/Ethnicity (% White)	55.6%		57.8%		0.51
Poverty:Income Ratio (Mean, SD)	2.34	1.55	2.38	1.59	0.16
BMI (Mean, SD)	27.01	6.84	26.61	6.8	0.08

Because the analytes rarely occur as single chemicals, isolated from other analytes, there is an inherent correlation among them, and similarly for the nutrients. The correlation patterns are summarized through the histograms of correlations in Figure 4.1. More complex higher degree relationships may still be present and are not indicated by pairwise correlations (Kutner, 2005).

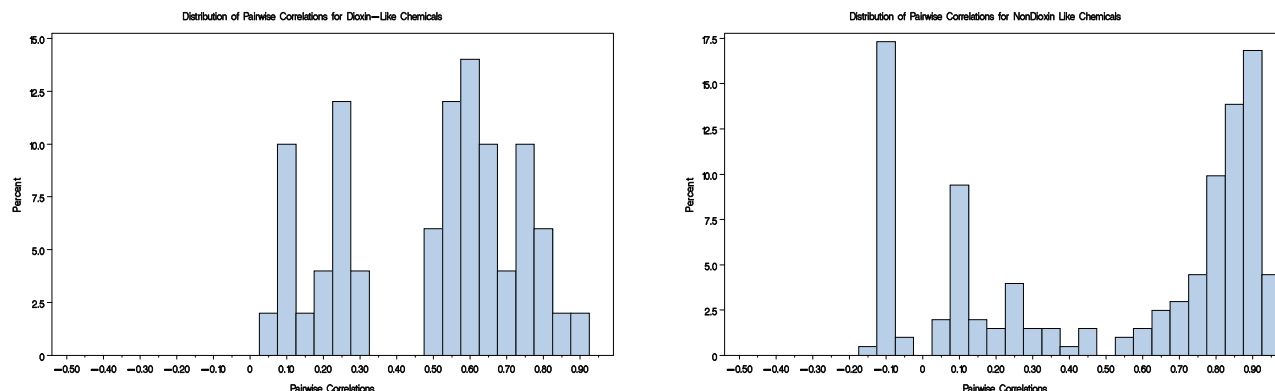
Figure 4.1: Correlations Among Environmental Chemicals and Nutrients (NOTE: Only those significantly different from zero are displayed. 100/666(15%) were nonsignificant for Chemicals and 386/1596(24%) for nutrients)



The histograms in Figure 4.1 show the complexity of the correlations among these two groups of components. The average absolute correlation (i.e. disregarding sign and only counting those significantly different from zero) for the chemicals is 0.48 with a range of 0.07 to 0.96. For the nutrients the average absolute correlation is 0.25, with a range of 0.06 to 0.97. Among the analytes and the dietary nutrients, there are logical explanations for both the high and low correlation values. Among the chemicals, the pairwise correlations among the co-planar PCBs are higher than those between a co-planar and a noncoplanar PCB. Figure 4.2 contains the pairwise correlations for dioxin-like and non-dioxin-like components, and comparing to Figure

4.1, the medium correlations (those around 0.4) are those that are missing. This implies that the pairwise correlations for a given dioxin-like PCB and a given non-dioxin-like PCB are in that low to medium range.

Figure 4.2: Distributions of Pairwise Correlations for Dioxin-Like and Non-Dioxin-Like



Similarly there are pairs of vitamins and minerals that occur together commonly and therefore have higher correlations. The set of B-vitamins (thiamin, riboflavin, niacin, folate, B6, B12) also have high pairwise correlations (all significantly different from 0 and most greater than 0.4). Low correlations can also be explained for the nutrients. For example, Vitamin C is primarily found in citrus fruits and B12 primarily comes from animal sources (meat, fish, eggs, milk); the two had an observed correlation of 0.02 which was not significantly different from 0.

These relationships contribute to the complex correlation structure among the analytes and the nutrients. Rather than reduce the dimensionality by considering only one chemical and one nutrient, our proposed weighted quartile score approach reduces the dimensionality without over simplifying the relationship between environmental chemicals, nutrients, and ALT.

Upon investigation of the continuous covariates, there was indication of a quadratic relationship between age and ALT. Therefore, the core model included the following covariates:

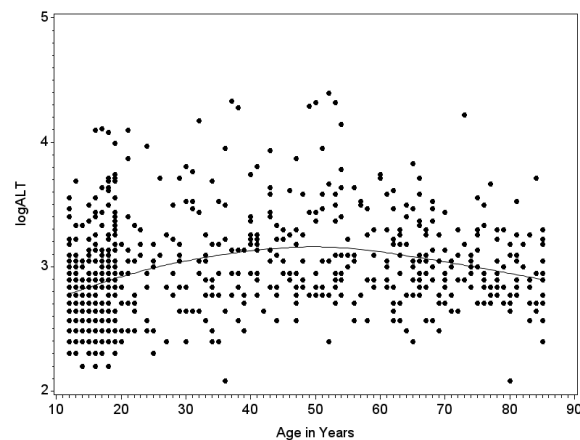
age (in decades), age² (with age in decades), gender, a binary race variable (Non-Hispanic White vs Others), BMI, and Poverty:Income ratio. The results of the core model analysis are given in Table 4.3 and a plot of age vs ALT is given in Figure 4.3. Although Poverty:Income ratio was not statistically significant, it was left in the model based on its importance in the literature (Christensen 2013, Cave 2010). No other continuous covariates had a quadratic or other higher order relationship with ALT.

Table 4.3: Core Model Assessment

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	P-Value
Intercept	2.47	0.074	2.262	2.554	1046.41	<.001
Age	0.241	0.037	0.168	0.314	41.80	<.001
Age*Age	-0.024	0.004	-0.032	-0.016	35.76	<.001
Race/Ethnicity (Other vs NH White)	0.067	0.035	-0.001	0.135	3.69	0.055
RIAGENDR (Female vs. Male)	-0.218	0.031	-0.280	-0.156	48.09	<.001
Poverty:Income Ratio	0.004	0.011	-0.017	0.025	0.15	0.702
BMI	0.008	0.003	0.003	0.013	8.85	0.003

*Age in decades

Figure 4.3: Plot of Age vs ALT to Demonstrate Quadratic relationship



3.2 Weighted Index for Environmental Chemicals Score (ECS)

Following the methods of Carrico, et al, we performed a bootstrap analysis to estimate average weights to define ECS. We took 1000 bootstrap samples of size 464 from the test dataset. The weights from each bootstrap were validated by fitting the model in 2.1 in each bootstrap sample. That is, the weights were estimated from the bootstrap sample data then the weights were used to construct the index. The weights were deemed significant if B1 was found to be significantly different from zero. The bootstrap sample weights were significant in 901 (of 1000) samples. The average weights from these samples were calculated, used to define ECS, and are given in Table 4.4. Those in bold have weights greater than 0.05.

Table 4.4 Average Bootstrap Weights for Environmental Chemicals

Dioxin-like compounds		Non-dioxin-like PCBs				Metals	
PCB 28	0.019	PCB 44	0.000	PCB 153	0.001	Cadmium	0.008
PCB 66	0.002	PCB 49	0.002	PCB 170	0.000	Lead	0.005
PCB 74	0.000	PCB 52	0.026	PCB 177	0.002	Mercury	0.106
PCB 105	0.001	PCB 87	0.017	PCB 178	0.002		
PCB 118	0.002	PCB 99	0.012	PCB 180	0.003		
PCB 156	0.001	PCB 101	0.128	PCB 183	0.010		
1,2,3, 6,7,8-HxCDD	0.001	PCB 110	0.016	PCB 187	0.007		
1,2,3,4,6,7,8-HPCDD	0.420	PCB 138	0.000	PCB 194	0.005		
1,2,3,4,6,7,8,9-OCDD	0.022	PCB 146	0.001	PCB 196	0.002		
1,2,3,4,6,7,8-HPCDF	0.000	PCB 149	0.000	PCB 206	0.002		
3,3',4,4',5-PNCB	0.173	PCB 151	0.004	PCB 209	0.001		
Dioxin-Like Total:	0.641	Non-Dioxin-Like Total:		0.241		Metals Total:	0.119

A majority of weight is placed on the dioxin-like compounds (64%). A total of 24% of the weight was on non-dioxin-like components, with just under 13% of the weight placed on PCB 101, and no other components with weight greater than 0.05. The three metals accounted for 12% of the weight, with just under 11% attributed to mercury. The two components with the highest weight are 1,2,3,4,6,7,8-HPCDD and 3,3',4,4',5-PNCB. Using the average weights in

Table 4.4 to define ECS, ECS was significant in the validation dataset; results are given in Table 4.5.

Table 4.5 Model Results for Average Bootstrap ECS

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald	Pvalue
Intercept	2.382	0.074	2.236	2.527	1030.91	<.001
ECS	0.103	0.034	0.037	0.169	9.40	0.002
Age	0.185	0.041	0.105	0.266	20.37	<.001
Age*Age	-0.021	0.004	-0.029	-0.013	25.31	<.001
Race (Others vs NH White)	0.086	0.035	0.017	0.154	5.99	0.014
Gender (Female vs Male)	-0.217	0.031	-0.278	-0.156	48.64	<.001
Poverty:Income Ratio	-0.001	0.011	-0.021	0.020	0.00	0.956
BMI	0.008	0.003	0.003	0.013	10.12	0.002

*Age in decades

These results indicate that for every one unit in the ECS, there is an average increase in $\log(\text{ALT})$ of 0.10, or an increase in mean ALT of 11% (i.e. $e^{0.10}=1.11$). In terms of the covariates, Non-hispanic whites are associated with a decrease in ALT over other races; Males are associated with an increase in ALT over Females. There was a quadratic affect for age which indicates that ALT increases with age until a certain point and then it begins to decrease again. There was no significant increase in ALT based on poverty:income ratio. Individuals with higher BMI were also associated with a modest increase in ALT.

3.3 Weighted Index for Nutrient Stressor Score (NSS)

In a similar manner, the 1000 bootstrap samples were used to estimate the average weights to define NSS. The average from the 1000 weights that were significant in their sample replicates are given in Table 4.6 (those greater than 0.05 are bolded in the table) and the parameter estimates for the model are given in Table 4.7. Since the estimate for NSS is positive, the weights indicate nutrients identified as non-chemical stressors. The vitamins and minerals

accounted for 37% of the total weights, with most weights less than 0.05, and three right around 0.05. The set of all types of fats accounted for 29% of the index with highest weight on MFA 22:1 (Docosenoic; an Omega-9 fatty acid). The remaining 33% of the weight was placed on five of the other seven components including: carbohydrates, fiber, protein and sugar.

Table 4.6 Nutrient Weights from Bootstrap Analysis

Vitamins/Minerals	Weight	Fats	Weight	Others	Weight
Vitamin E	0.002	Total Fat	0.000	Protein	0.034
Vitamin A	0.006	Total Saturated Fatty Acids (SFA)	0.028	Carbohydrates	0.098
Alpha Carotene	0.022	SFA 4:0 (Butanoic)	0.012	Total Sugars	0.078
Beta Carotene	0.041	SFA 6:0 (Hexanoic)	0.004	Dietary Fiber	0.082
Beta Cryptoxanthin	0.024	SFA 8:0 (Octanoic)	0.000	Sodium	0.015
Lycopene	0.022	SFA 10:0 (Decanoic)	0.002	Caffeine	0.018
Lutein+zeaxanthin	0.009	SFA 12:0 (Dodecanoic)	0.001	Theobromine	0.008
Thiamin	0.026	SFA 14:0 (Tetradecanoic)	0.000		
Riboflavin	0.000	SFA 16:0 (Hexadecanoic)	0.012		
Niacin	0.003	SFA 18:0 (Octadecanoic)	0.002		
Vitamin B6	0.001	Total Monounsaturated Fatty Acids (MFA)	0.000		
Folate	0.002	MFA 16:1 (Hexadecenoic)	0.006		
Folic Acid	0.003	MFA 18:1 (Octadecenoic)	0.018		
Food Folate	0.055	MFA 20:1 (Eicosenoic)	0.038		
Vitamin B12	0.002	MFA 22:1 (Docosenoic)	0.060		
Vitamin C	0.014	Total Polyunsaturated Fatty Acids (PFA)	0.003		
Vitamin K	0.008	PFA 18:2 (Octadecadienoic)	0.011		
Calcium	0.004	PFA 18:3 (Octadecatrienoic)	0.004		
Phosphorus	0.050	PFA 20:4 (Eicosatetraenoic)	0.006		
Magnesium	0.003	PFA 20:5 (Eicosapentaenoic)	0.030		
Iron	0.001	PFA 22:5 (Docosapentaenoic)	0.001		
Zinc	0.001	PFA 22:6 (Docosahexaenoic)	0.051		
Copper	0.051				
Potassium	0.002				
Selenium	0.022				
Vit/Min Total:	0.37	Fats total:	0.29	Others Total:	0.33

Table 4.7: Model Estimation with Average Bootstrap Weights for Nutrient Non-Chemical Stressors

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Wald	Pvalue
Intercept	2.326	0.079	2.172	2.481	868.29	<.001
NSS	0.123	0.044	0.037	0.210	7.79	0.005
Age	0.206	0.032	0.143	0.270	40.73	<.001
Age*Age	-0.022	0.003	-0.029	-0.015	40.57	<.001
Race (Others vs NH White)	0.059	0.029	0.001	0.116	4.00	0.045
Gender (Female vs Male)	-0.248	0.026	-0.299	-0.197	89.64	<.001
Poverty:Income Ratio	0.003	0.009	-0.015	0.020	0.08	0.774
BMI	0.008	0.002	0.004	0.012	13.54	<.001

*Age in decades

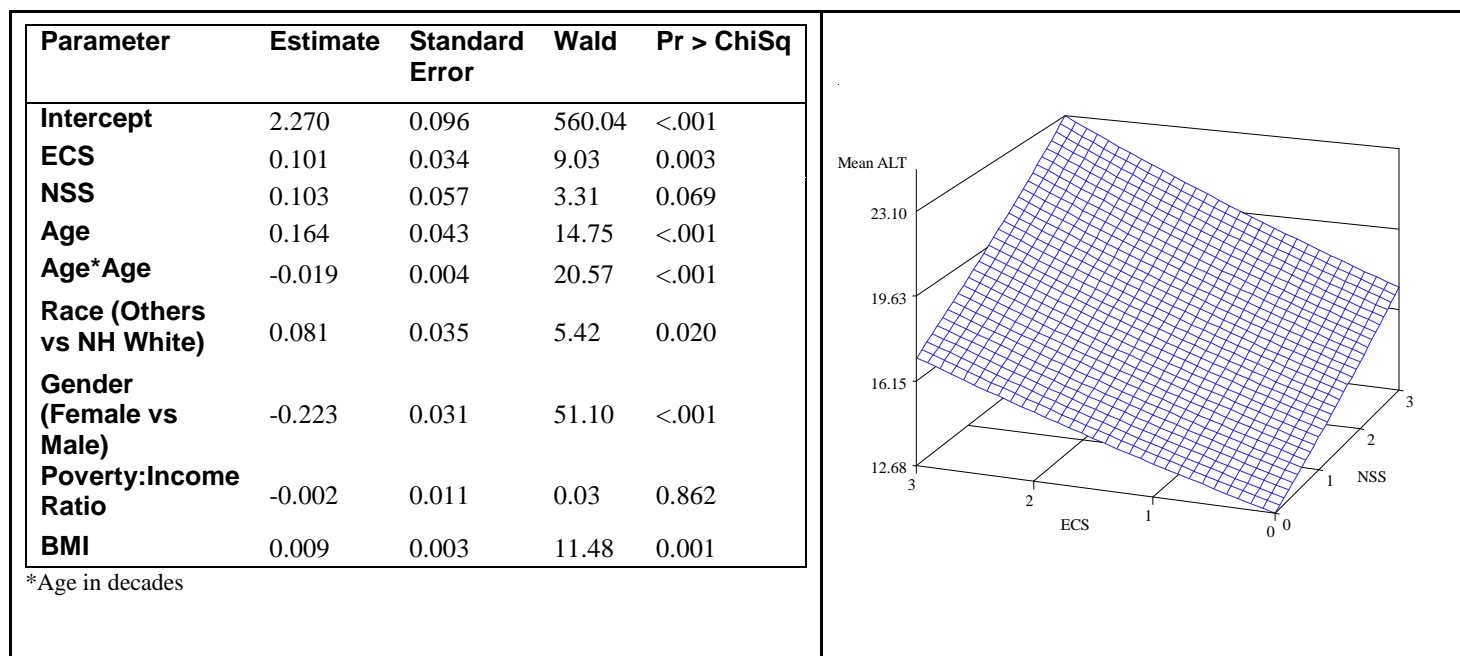
The NSS was significant in the validation dataset. Based on these results, a one unit increase in NSS is associated with a 0.12 increase in mean log(ALT), or 13% increase in mean ALT. Again age had a significant quadratic relationship with ALT; females were associated with a lower ALT; BMI was associated with an increase in ALT; Poverty:income ratio was not significant. In this model, Whites were associated with a decrease in ALT as compared to other races. In the model with ECS, this relationship was opposite, however it was not statistically significant in the prior model.

3.4: Joint Model for Chemicals and Nutrients

Using ECS and NSS, we fit the model in 2.3d and found that the interaction term was not statistically significant (p-value=0.32). Following the methods, the model in 2.3c was used as the final model. The results are given in Figure 4.4, along with a figure of the predicted mean ALT at the average level of the covariates. The significance and direction of the covariates are similar to the covariates in the previous two models. Both ECS and NSS are positive. ECS was significant (p=0.003) and NSS was marginally significant (p-value=0.069) in the independent

validation dataset. From the figure it can be seen that the highest levels of ALT are predicted for individuals with high levels of ECS and high NSS.

Figure 4.4: Model Estimation With Chemical and Nutrient Index and Predicted Mean ALT for ECS and NSS at Average Levels for BMI, Gender, Race (binary), PIR, and Age



Discussion

4.1 Implication of Indices

Considering ECS, a majority of the weights were assigned to the dioxin-like chemicals (dioxins, furans and coplanar PCBs) and the heavy metals, with almost half of the total weight assigned to 1,2,3,4,6,7,8-HPCDD. Cave, et al (2010) stated that coplanar PCBs and mercury both concentrate primarily within the liver, while noncoplanar PCBs concentrate in the adipose tissue which could be an explanation for their associations with an increase in ALT (Klein 1972; Mudipalli 2007; National Toxicology Program (NTP) 2006). If coplanar PCBs and

mercury are concentrated within the liver it is likely that they could cause greater stress.

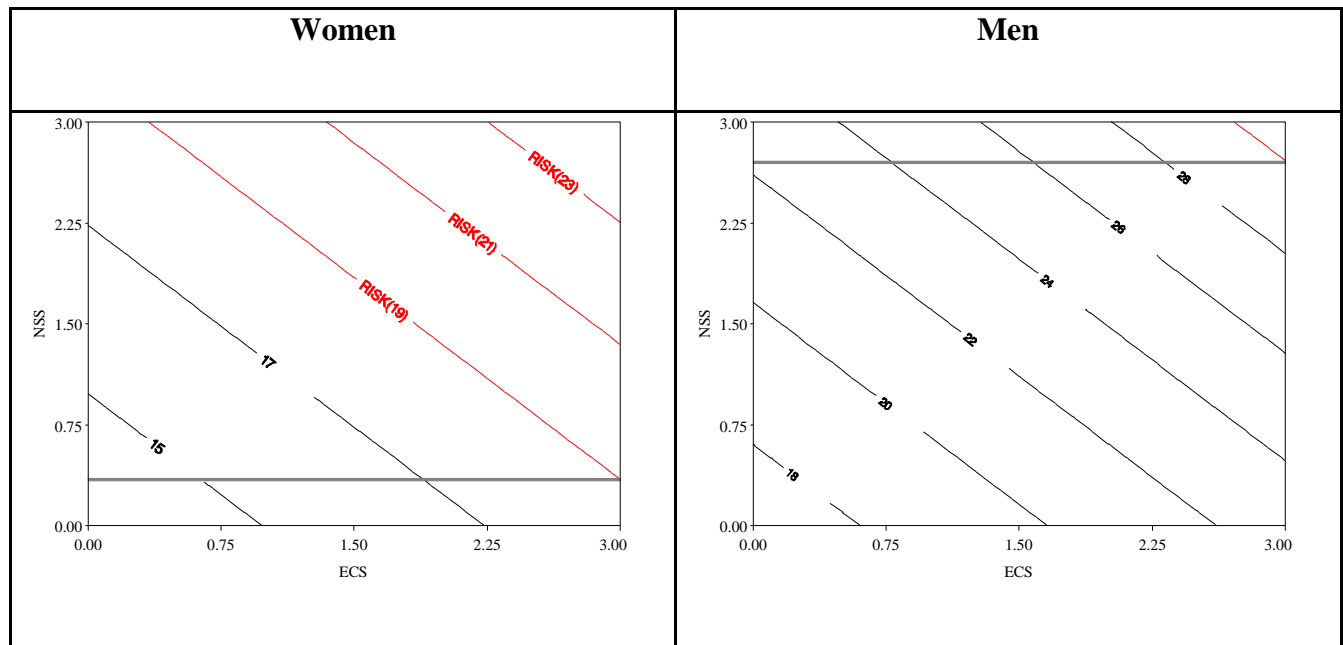
Additionally, in animal studies, male mice were exposed to drinking water with a mixture of mercury, lead, cadmium and copper saw an increase in ALT (Al-Attar, 2011).

Considering NSS, since all components were adjusted to total caloric intake, the high weight on a component can be interpreted as a association between ALT and a high percentage of calories coming from a given source (i.e. for caffeine if a high percentage of calories come from drinks with high caffeine content are indicative of increase in ALT). Carbohydrates, sugars and fiber were the main components assigned weights greater than 0.05. York (2009) defines NAFLD as a two-hit process, the first of which likely being caused by insulin resistance. Because sugars and carbohydrates are converted to glucose in the body and most individuals with fatty liver disease have insulin resistance (York 2009), having a diet high in calories from carbohydrates (sugar and sources of fiber) leads to higher glucose production and without the necessary insulin production the excess glucose is converted to fat and can be deposited onto the liver (Nordlie, et al 1999; Wilcox 2005). Because the liver is responsible for the metabolism of protein, having a diet high in calories from protein may place excess stress on the liver. For this reason, individuals with liver diseases and disorders are often recommended to limit protein intake (Academy of Nutrition and Dietetics).

Clinical thresholds for ALT are 19u/L for women and 30u/L for men (Assy, 2009). Based on the model for average ALT, we can predict what levels of ECS and NSS will help a person achieve a healthy ALT level. However, for almost all people, the ECS value is unknown. What is known is that it is certainly greater than zero based on limit of detection values from NHANES data. For that reason, we will consider the contour figures in 4.5. The figure for men indicates that without knowledge of chemical exposure, a man (on average- with average levels of the

covariates) needs to have a NSS of less than 2.7, while a woman would need to have a score at most 0.34 assuming the worst chemical exposure score. So in order to reduce liver stress, men are allowed more fluctuation in their diets than women.

Figure 4.5: Contour Plots for Men and Women Average ALT Versus ECS and NSS



The weighted index for the nutrients also allows for dietary recommendations for individuals concerned about liver health. From the analysis, we can say that a diet rich in fruits and vegetables is better for liver health than one high in carbohydrates, proteins and fats. The Academy of Nutrition and Dietetics also suggests that individuals with symptoms of liver disease (including NASH/NAFLD) should avoid excess sodium, fluid, fats and sugars. This is in line with the form of NSS, with the lowest weight on vitamins and minerals (i.e. fruits and vegetables) and higher weights on sugars, carbohydrates, and proteins. For both men and women to have a healthy NSS level and therefore reduce the stress on the liver, they should strive for a diet rich in fruits and vegetables, with limited protein, carbohydrates and minimal caffeine.

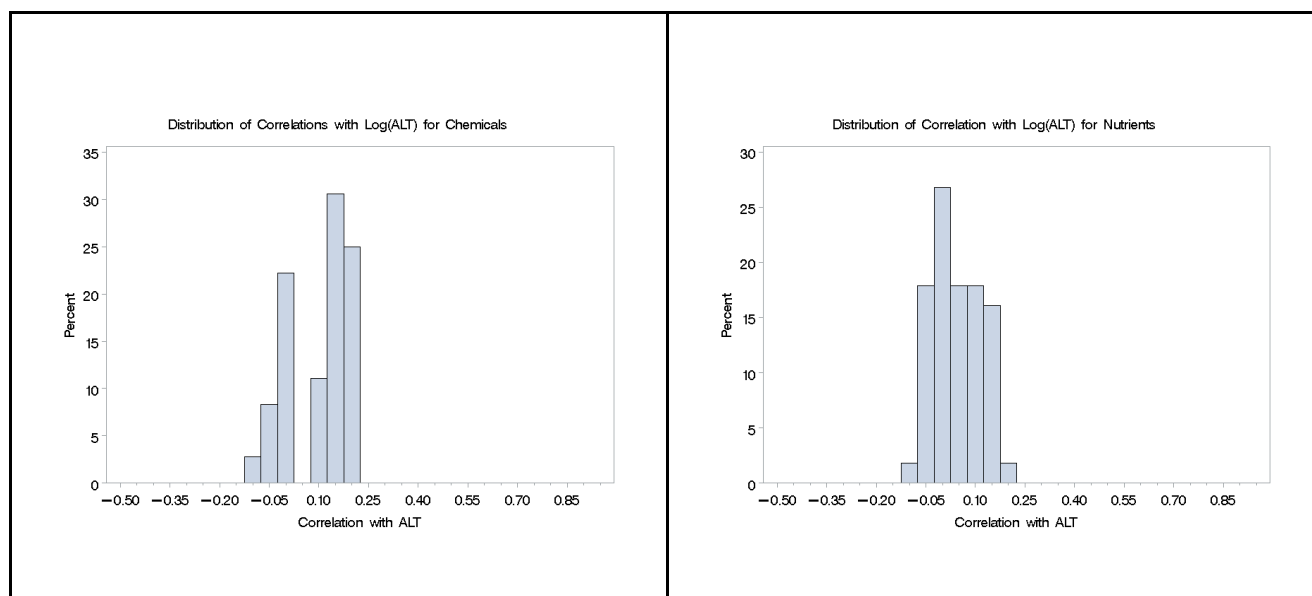
4.2 Limitations

Because NHANES is a cross-sectional study, only associations can be suggested. No causal relationships can be determined or suggested. There also may be other nutrients and environmental chemicals that were not considered by this analysis that are associated with increased ALT. We also did not estimate an index that identifies nutrients that are “protective” rather than non-chemical stressors. Because the weights used to calculate the average bootstrap index are those that are significant in the bootstrap samples, they were not independently validated. However, this allows for the validation dataset to remain independent allowing for a true test of significance.

Despite limitations, this method allows for multiple chemicals and multiple nutrients (both with complex correlation structures) to be analyzed in the same model despite complex correlation structures. The simulations in Chapter 2 characterized the weighted quantile score approach and demonstrated that in situations with high pairwise correlations among the components and low correlations with the outcome, a bootstrap analysis will lead to improved estimates for the weights. We concluded that among sets of highly correlated components, weights may be diminished due to the high correlations but that a zero weight is indicative of a lack of association with an outcome and that a nonzero weight is indicative of an association with the outcome variable. In our case, the pairwise correlations are in many cases greater than 0.9 and overall very complex. Additionally, the correlations with the outcome are not extremely high (See Figure 4.6). This case is likely a “breakdown case” as described in Chapter 2, but the additional bootstrap analysis was shown to offer improvement over a single estimation of weights. The simulations in Chapter 2 demonstrated that the addition of the bootstrap analysis improves accuracy through improved reliability and validity. The small weights (roughly 0.01-

0.05) seen in both the chemical and the nutrients reflect the complex correlations among the chemicals and nutrients and not lack of association (i.e. 0 weights). Both ECS and NSS can also be interpreted as indices indicating associations between persistent environmental chemicals, poor nutrition and liver toxicity as measured by ALT.

Figure 4.6: Distribution of Correlations with Log(ALT) for Chemicals and Nutrients



V. Conclusions and Future Work

5.1 Conclusions

The goal of this thesis was to characterize and apply a weighted quantile approach for a risk analysis setting. This method is motivated by complex correlation patterns in environmental chemicals or other groups of potential predictors with high correlations and potential multicollinearity problems. In order to characterize the method, a heuristic argument was made and extensive simulations were performed. To demonstrate the usefulness of the method, we presented an example using a weighted quantile score approach that modeled liver health (ALT) as it relates to environmental chemicals and nutrition.

The heuristic argument argued that the addition of the constraint on the weights (they sum to 1) stabilizes the optimization process by decreasing the eigenvalue spectrum of the hessian. Through simulations, it was shown that the weighted quantile score approach performs better than ordinary least squares and LASSO. We also demonstrated that in settings where the pairwise correlations are smaller compared to the correlations with the outcome, there is breakdown. However, in the breakdown cases (low sample size, high pairwise correlations and low correlations with the outcome), the addition of a bootstrap analysis and calculating average weights, all important components can be detected. However, we have seen that components with high pairwise correlations may have slightly smaller weights. That is, for the same correlation with Y , components with high pairwise correlations will have smaller weights than a component with the same correlation with the outcome which is independent (or less correlated). In cases where components are independent, weights can be interpreted as association with the outcome relative to the other components. In cases with complex correlation patterns, weights are influenced by both importance with the outcome and the correlation structure. The

simulations demonstrate that the average bootstrap weights for components with no correlation with the outcome will be zero or “very near” zero. This is a benefit of the weighted quantile score over ordinary regression and LASSO.

After characterizing the approach, we applied and interpreted results using real data. NHANES biomonitoring data for blood serum levels of PCBs, dioxins, furans and heavy metals and 2-day total dietary recall were used to estimate a score for environmental chemicals and a score for nutrition as a non-chemical stressor. The correlation pattern for both the environmental chemicals and the nutrients was extremely complex including several pairwise correlations that were near perfect (greater than 0.90). The average weighted score for both the environmental chemicals and the nutritional data were found to be significantly associated with increase in ALT in a validation dataset. Because the weights indicated nutritional components associated with an increase in ALT, high levels of these nutrients may be considered non-chemical stressors. There was no significant interaction between the index for the environmental chemicals and nutritional status. Because the weighted quantile score approach can be used in a setting with complex correlation structure, we were able to model a large number of chemicals and nutrients and their effect on a health outcome like ALT. With the weighted quantile approach, a more complete assessment of the relationship between these two stressors have on ALT can be ascertained visualized.

5.2 Future Work

The method at hand is defined as a weighted quantile score approach, with all examples and illustrations done with quartile scoring. The determination for the appropriate number of quantiles that should be used has not been evaluated. Quartiles were used because they are the most common in the current literature (e.g. Swan, et al; Cave, et al 2010). Future work could

entail developing methods to determine the optimal number of quantiles the components of the score should be divided into. Additionally, each component could be scored into a different number of quantiles.

The quartile scoring performed in this thesis also assumes a linear relationship between the components and the outcome. However, there could be a potential quadratic effect. For example, it could be true that protein (which was seen to increase ALT in Chapter 4) could actually be beneficial to health up to a certain point at which the relationship becomes negative (i.e. leads to an increase in ALT). It seems reasonable to imagine that a certain amount of protein is fine but that there may be a limit to the amount that the liver can metabolize without causing undue stress and damage. For this reason, a potential quadratic effect of the quantile-scored components could be considered.

While we suggest that an increased sample size will have an effect on the stability of the results, we do not develop a rule-of-thumb or suggestion for how large of a sample size is needed for the analysis. We also only use an equal split for the test/validation datasets. There is debate about whether a larger sample size should be used to estimate the weights (higher stability) or if a larger sample size should be used for the validation (i.e. higher power). There is certainly a call for research on whether or not a 40/60, 60/40, or any other potential split of the data could be more optimal.

The weights here were all chosen based on the optimization of the likelihood. In the examples in this dissertation, log-linear models were discussed, but this certainly can be expanded to any likelihood. A paper by Gennings, et al (2013) uses the weighted quantile score approach to model time-to-pregnancy using a Weibull survival model. Additionally, a different optimization criterion could be considered. If the goal is to estimate and predict an outcome, then

a prediction error objective function could be used rather than the maximum likelihood objective function.

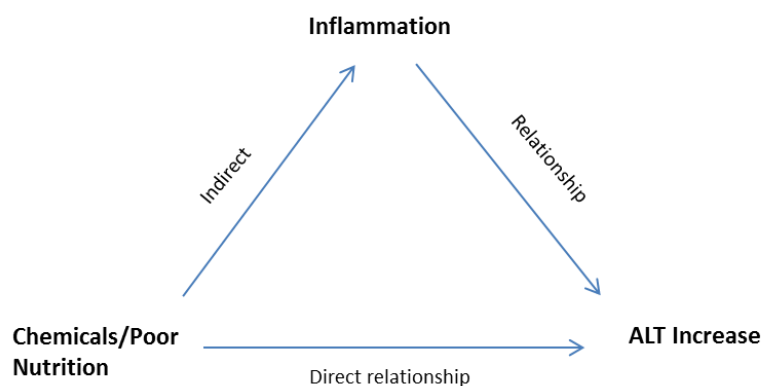
It is clear that negative environmental chemical exposures cannot be completely alleviated and that their presence and effects are widespread. We see levels of many chemicals at values above the limit of detection in virtually all subjects in the NHANES datasets. Despite the ban of many environmental chemicals (i.e. PCBs), due to their persistent nature, levels of these are still detected at high rates in populations. For this reason, there is an interest in risk analysis to determine if there are possible mitigating or protective effects like good nutrition that can diminish the negative effects of environmental chemical exposures. In a mathematical sense, this would be an interaction between two indices: one for the negative environmental chemicals and possibly non-chemical stressors and one for these potential protective components. This method could include an interaction term in the model and even potentially optimize the weights based on the interaction term. That is, the optimization could be set to determine what weights optimize the parameter for the interaction. This can be thought of like a ds-optimal design which finds a study design that minimizes the variance of a subset of the parameters. We could have the optimization criterion be related to the precision of the parameter estimation for the interaction term. A more basic approach would be to include the interaction term in the likelihood and optimize the same as before. However, with indices with large number of components, this may lead to difficulty in estimation and a need for a larger sample size.

We also assume that there is a direct relationship between the components of the index and the health outcome chosen. We do not consider potential mediators in the system.

Considering the application in Chapter 4: What if the environmental chemicals or non-chemical stressor nutrients actually cause insulin resistance and that causes the increase in ALT? If an

index (weighted quantile score) and a mediator are placed in the same model, the effect of the index may decrease. Consider the diagram in Figure 5.1. There may be little possibility to distinguish between the two situations, especially with cross-sectional data like that in NHANES.

Figure 5.1 Diagram of Possible Mediating Effect



Conclusion

While there are certainly limitations to the weighted quantile approach as presented here, the method has been shown to be a promising option in a risk analysis setting. The possible extensions to this work will add to the importance and usefulness of this method.

The weighted quantile score method is developed for a risk analysis setting where the goal is the identification of “bad actors.” We have shown that the WQS has good validity and reliability, especially in cases with higher correlation with the outcome compared to the pairwise correlations. In some cases, the method has also demonstrated stability and benefits over current methods like OR and LASSO due to the reduction in both false positive and false negative rates. While we do not propose the method be used in all modeling, it is a good option for modeling highly correlated data when there is a logical grouping (chemicals, nutrients, etc).

Appendix 1: References

References

- Al-Attar, A. (2011). Vitamin E attenuates liver injury induced by exposure to lead, mercury, cadmium and copper in albino mice. *Saudi Journal of Biological Sciences*, 18(4), 395-401.
- Anderson, W. and Wells, M. (2008). Numerical Analysis in Least Squares Regression with an Application to the Abortion-Crime Debate. *Journal of Empirical Legal Studies*, 5: 647–681.
- Assy N., Beniashvili Z., Djibre A., Nasser G., Grosovski M., Nseir W. (2009). Lower baseline ALT cut-off values and HBV DNA levels better differentiate HBeAg(-) chronic hepatitis B patients from inactive chronic carriers. *World Journal of Gastroenterology*, 15(24), 3025.
- Cave, M., Deaciuc, I., Mendez, C., Song, Z., Joshi-Barve, S., Barve, S., et al (2007). Nonalcoholic fatty liver disease: Predisposing factors and the role of nutrition. *The Journal of Nutritional Biochemistry*, 18(3), 184-195.
- Cave, M., Appana, S., Patel, M., Keith Cameron Falkner, McClain, C. J., & Brock, G. (2010). Polychlorinated biphenyls, lead, and mercury are associated with liver disease in American adults: NHANES 2003-2004. *Environmental Health Perspectives*, 118(12), 1735-1742.
- Center for Disease Control. National Health and Nutrition Examination Study.
<http://www.cdc.gov/nchs/nhanes.htm>.
- Gennings, C., Sabo, R., & Carney, E. (2010). Identifying subsets of complex mixtures most associated with complex diseases polychlorinated biphenyls and endometriosis as a case study. *Epidemiology*, 21, S77-S84.

- Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. Dordrecht: Dordrecht Springer-Verlag New York Inc.
- Hennig, B., Oesterling, E., & Toborek, M. (2007). Environmental toxicity, nutrition, and gene interactions in the development of atherosclerosis. *Nutrition, Metabolism, and Cardiovascular Diseases : NMCD*, 17(2), 162-169.
- Krawczyk, M., Bonfrate, L., Portincasa, P. (2010). Nonalcoholic fatty liver disease. *Best Practice & Research. Clinical Gastroenterology*, 24(5), 695.
- Krigman, M. R., & Krigman, M. R. (1972). *A model of acute methyl mercury intoxication in rats; archives of pathology*
- Kutner, M. H. (2005). *Applied linear statistical models michael H. kutner ... [et al.]* (5th ed.). Boston: McGraw-Hill Irwin.
- Lewis, A. S., Sax, S. N., Wason, S. C., Campleman, S. L. (2011). Non-chemical stressors and cumulative risk assessment: An overview of current initiatives and potential air pollutant interactions. *International Journal of Environmental Research and Public Health*, 8(6), 2020-2073.
- Mudipalli A. (2007). Lead hepatotoxicity and potential health effects. *Indian J Med Res*. 126((6)):518–527.
- National Digestive Diseases Information Clearinghouse. *Nonalcoholic Steatohepatitis*.
<http://www.digestive.niddk.nih.gov/ddiseases/pubs/nash/>
- Nocedal, J. (2006). *Numerical optimization*. New York: New York : Springer.

Nordlie, R. C., Foster, J. D., Lange, A. J. (1999). Regulation of glucose production by the liver.

Annual Review of Nutrition, Vol.19(1), p.379-406

NTP (National Toxicology Program) (2006). NTP technical report on the toxicology and carcinogenesis studies of 2,2',4,4',5,5'-hexachlorobiphenyl (PCB 153) (CAS no. 35065-27-1) in female Harlan Sprague-Dawley rats (gavage studies) Natl Toxicol Program Tech Rep Ser. (529):4-168

SAS Institute Inc (2008). *SAS 9.2 Help and Documentation*. Cary, NC: SAS Institute Inc.

Swan, S. H. (2008). Environmental phthalate exposure in relation to reproductive outcomes and other health endpoints in humans. *Environmental Research*, 108(2), 177-184.

Wilcox, G. (2005) Insulin and insulin resistance. *The Clinical Biochemist.Reviews / Australian Association of Clinical Biochemists*, 26(2), 19.

York, L. W., Puthalapattu, S., & Wu, G. Y. (2009). Nonalcoholic fatty liver disease and low-carbohydrate diets. *Annual Review of Nutrition*, 29, 365-379.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society.Series B (Statistical Methodology)*, 67(2), 301-320.

Appendix II: SAS Code

A2.1. Simulating Correlated Data

```
proc iml;

    varnames = { /*Column headings in quotes; space delimited for example: 'y' 'x1'
'x2'... */ };
    varnum = ncol(varnames);
    mean = 1.5#j(varnum,1,1); /*Note 1.5 is mean for quartile scored data (0-3;
mean[1]=51; /*Input Mean for outcome variable;
se = 1.1#j(varnum,1,1); /*Note: 1.1 is Std Error for quartile scored data
se[1]= 15; /*SE for Y;
    number = 2000; /*Number of observations in TOTAL dataset

    corr={ /*Input Correlation Matrix: space delimited correlation values; comma to
separate rows */ };

    ridge=j(1,12, /*input ridge value; must be greater than 1 as this will be the value for the
diagonal */ );
    corr= (corr-i(12))+diag(ridge); /*subtracts off the 1's on the diagonal and replaces with
ridge value;
    var = diag(se)*corr*diag(se);
    chol = half(var);

    do sample = 1 to 1000; /*change 1000 to the number of simulations desired;
    call randseed(12345);

    /* get number random observations from standard normal */
    yx = j(number,varnum,.);
    /* each row of m comes from a different distribution */
    call randgen(yx,'NORMAL'); /** standard normal;

    yx = j(number,1,1)*mean` + yx*chol;

    percent=0.5; /*indicates a 50-50 split of the data for test/validation datasets;
any percent value between 0 and 1 can be used and will determine the split of the
test/validate datasets;

    results = (sample#j(number,1,1))|| yx || (j(percent#number,1,1)/j((1-
percent)#number,1,2) );

    reslabels = 'sample' || varnames || 'group';
    all = all // results;
end;

create all from all[colname=reslabels]; append from all; run;
```

A2.2 Macro for Simulating Bootstrap Samples

```
proc iml;
  varnames = {/*INPUT VARIABLE HEADINGS*/};
  varnum = ncol(varnames);
  mean = 1.5#j(varnum,1,1);
  mean[1]=51;
  se = 1.1#j(varnum,1,1);
  se[1]= 15;
  number = 500;
  ridge=j(1,12,1); /*CHANGE RIDGE*/
  corr= (corr-i(12))+diag(ridge);
  var = diag(se)*corr*diag(se);
  chol = half(var);

  do sample = 1 to 1000; /*INPUT THE NUMBER OF SAMPLES
  DESIRED*/
  call randseed(12345678);

  /* get number random observations from standard normal */
  yx = j(number,varnum,.);
  /* each row of m comes from a different distribution */
  call randgen(yx,'NORMAL'); ** standard normal;
  yx = j(number,1,1)*mean` + yx*chol;
  percent=0.5;
  results = (sample#j(number,1,1))|| yx || (j(percent#number,1,1)/j((1-
  percent)#number,1,2) );
  reslabels = 'sample' || varnames || 'group';
  all = all // results;
  end;
  create all from all[colname=reslabels]; append from all;
  run;

  /*RANK SIMULATED DATA BY SAMPLE*/
  proc rank data=all groups=4 out=ranked ;
    by sample;
    var y x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 ;
    ranks yq x1q x2q x3q x4q x5q x6q x7q x8q x9q x10q x11q;
  run;
```

```

%macro sim(T);
    %do i = 1 %to &T;

data data;
set ranked;
if sample=&i;
run;

/*RANDOMLY SELECT NUMBER OF DESIRED BOOTSTRAP SAMPLES (e.g.
reps=100)*/
proc surveysselect data=data method=urs n=250 /*n=number of observations used to
estimate the weights; i.e. the size of the test dataset*/
reps=100 seed=113084 outhits out=test.bootstrap;      /*(reps=number of bootstrap
samples*/
strata sample;
run;

data test.start;
    _type_='PARMS';

/*Define starting values for parameters in model- NOT including weights*/
alpha=-6; beta1=0.1; sigma=1.0;

/*uses only group=1, i.e. TEST dataset*/
group=1;

/*Input weights- starting value is usually 1/c, where c is the number of components*/
array inwts /*list of weights, space delimited*/;

do over inwts;

    inwts=1/(/*NUMBER OF COMPONENTS IN INDEX- i.e. starting value is
equal for all*/);
end;
run;
proc nlp data=test.bootstrap technique= trureg
    maxiter=10000 maxfunc=10000
    inest=test.start /*uses dataset for starting values*/
    outest=outstuff /*creates output dataset with weights*/
    noprint;

```

```

        by replicate; /*INDICATES THAT WEIGHTS BE ESTIMATED FOR
        EACH BOOTSTRAP SAMPLE*/
max logL;
parms /INPUT PARAMETERS FOR MODEL*/;

logL= /*INPUT LIKELIHOOD FOR MODEL*/;


lincon /*INPUT CONSTRAINT: wx1q + wx2q + wx3q + ...= 1;*/

bounds /*INPUT BOUNDS: 0<wx1q<1, 0<wx2q<1, ...*/


run;


/*Create dataset with weights for each SAMPLE*/
data weights&i;
    set outstuff;
    where _type_='PARMS';
    sample=&i;

run;


/*Merge Test Dataset and Weights so WQS can be tested*/
data test&i;
merge weights&i test.bootstrap; by replicate;
where sample=&i;
run;


/*Create Variable for Index based on weights for each bootstrap sample*/
data test&i;
set test&i;
wted_sum= wx1q*x1q + wx2q*x2q + wx3q*x3q + wx4q*x4q + wx5q*x5q +
wx6q*x6q
+ wx7q*x7q + wx8q*x8q + wx9q*x9q + wx10q*x10q + wx11q*x11q ;
run;


/*Fits Model- Other procedures can be used here depending on the desired model*/

```

```

proc reg data=test&i;
model y=wted_sum;
by replicate;
ods output ParameterEstimates=parms&i;
run;

/*Creates a variable in dataset to determine if the WQS is significant*/
data parms&i noprint;
set work.parms&i;
where variable = 'wted_sum';
if tvalue >=1.96 then power=1; else power=0;
sample=&i;
run;

%end;

/*MERGE All Weights Datasets together*/
data test.FILENAME; /*CHANGE FILE NAME*/
    set weights1-weights&T;
run;

/*MERGE all results from Model fits into 1 dataset*/
data test.FILENAME; /*CHANGE FILE NAME*/
    set parms1-parms&T;
run;

/*SORT Merged Data*/
proc sort data=test.FILENAME; /*CHANGE FILE NAME*/
by sample replicate;

/*SORT Merged Data*/
proc sort data=test.FILENAME; /*CHANGE FILE NAME*/
by sample replicate;
run;

/*Merge 2 Datasets Into One*/
data test.all; /*CHANGE FILE NAME*/
    merge test.FILENAME test.FILENAME; /*CHANGE FILE NAMES*/
    by sample replicate;
run;
%mend;

```


%*average*(DRTACAR,DR1TACAR,DR2TACAR);
 %*average*(DRTALCO,DR1TALCO,DR2TALCO);
 %*average*(DRTATOA,DR1TATOA,DR2TATOA);
 %*average*(DRTATOC,DR1TATOC,DR2TATOC);
 %*average*(DRTB12A,DR1TB12A,DR2TB12A);
 %*average*(DRTBCAR,DR1TBCAR,DR2TBCAR);
 %*average*(DRTCAFF,DR1TCAFF,DR2TCAFF);
 %*average*(DRTCALC,DR1TCALC,DR2TCALC);
 %*average*(DRTCARB,DR1TCARB,DR2TCARB);
 %*average*(DRTCHL,DR1TCHL,DR2TCHL);
 %*average*(DRTCHOL,DR1TCHOL,DR2TCHOL);
 %*average*(DRTCOPP,DR1TCOPP,DR2TCOPP);
 %*average*(DRTCRYP,DR1TCRYP,DR2TCRYP);
 %*average*(DRTFA,DR1TFA,DR2TFA);
 %*average*(DRTFDFE,DR1TFDFE,DR2TFDFE);
 %*average*(DRTFF,DR1TFF,DR2TFF);
 %*average*(DRTFIBE,DR1TFIBE,DR2TFIBE);
 %*average*(DRTFOLA,DR1TFOLA,DR2TFOLA);
 %*average*(DRTIRON,DR1TIRON,DR2TIRON);
 %*average*(DRTKCAL,DR1TKCAL,DR2TKCAL);
 %*average*(DRTLYCO,DR1TLYCO,DR2TLYCO);
 %*average*(DRTLZ,DR1TLZ,DR2TLZ);
 %*average*(DRTM161,DR1TM161,DR2TM161);
 %*average*(DRTM181,DR1TM181,DR2TM181);
 %*average*(DRTM201,DR1TM201,DR2TM201);
 %*average*(DRTM221,DR1TM221,DR2TM221);
 %*average*(DRTMAGN,DR1TMAGN,DR2TMAGN);
 %*average*(DRTMFAT,DR1TMFAT,DR2TMFAT);
 %*average*(DRTNIAC,DR1TNIAC,DR2TNIAC);
 %*average*(DRTP182,DR1TP182,DR2TP182);
 %*average*(DRTP183,DR1TP183,DR2TP183);
 %*average*(DRTP184,DR1TP184,DR2TP184);
 %*average*(DRTP204,DR1TP204,DR2TP204);
 %*average*(DRTP205,DR1TP205,DR2TP205);
 %*average*(DRTP225,DR1TP225,DR2TP225);
 %*average*(DRTP226,DR1TP226,DR2TP226);
 %*average*(DRTPFAT,DR1TPFAT,DR2TPFAT);
 %*average*(DRTPHOS,DR1TPHOS,DR2TPHOS);
 %*average*(DRTPOTA,DR1TPOTA,DR2TPOTA);
 %*average*(DRTPROT,DR1TPROT,DR2TPROT);
 %*average*(DRTRET,DR1TRET,DR2TRET);
 %*average*(DRTS040,DR1TS040,DR2TS040);
 %*average*(DRTS060,DR1TS060,DR2TS060);
 %*average*(DRTS080,DR1TS080,DR2TS080);
 %*average*(DRTS100,DR1TS100,DR2TS100);
 %*average*(DRTS120,DR1TS120,DR2TS120);
 %*average*(DRTS140,DR1TS140,DR2TS140);
 %*average*(DRTS160,DR1TS160,DR2TS160);

```

%average(DRTS180,DR1TS180,DR2TS180);
%average(DRTSELE,DR1TSELE,DR2TSELE);
%average(DRTSFAT,DR1TSFAT,DR2TSFAT);
%average(DRTSODI,DR1TSODI,DR2TSODI);
%average(DRTSUGR,DR1TSUGR,DR2TSUGR);
%average(DRTTFAT,DR1TTFAT,DR2TTFAT);
%average(DRTTHEO,DR1TTHEO,DR2TTHEO);
%average(DRTVARA,DR1TVARA,DR2TVARA);
%average(DRTVB1,DR1TVB1,DR2TVB1);
%average(DRTVB2,DR1TVB2,DR2TVB2);
%average(DRTVB6,DR1TVB6,DR2TVB6);
%average(DRTVB12,DR1TVB12,DR2TVB12);
%average(DRTVC,DR1TVC,DR2TVC);
%average(DRTVD,DR1TVD,DR2TVD);
%average(DRTVK,DR1TVK,DR2TVK);
%average(DRTZINC,DR1TZINC,DR2TZINC);

run;

proc contents data=app.pcb_nutr_data; run;

/*NO ALCO, ATOA, B12A, TP184*/
data app.pcb_nutr_data;
set app.pcb_nutr_data;
nmiss=nmiss( LBX028LA, LBX066LA, LBX074LA, LBX105LA, LBX118LA, LBX156LA,
             LBXD03LA, LBXD05LA, LBXD07LA, LBXF08LA, LBXPCBLA ,
             LBX044LA, LBX049LA, LBX052LA, LBX087LA, LBX099LA, LBX101LA,
LBX110LA, LBX138LA,
             LBX146LA, LBX149LA, LBX151LA, LBX153LA, LBX170LA, LBX177LA,
LBX178LA, LBX180LA, LBX183LA, LBX187LA, LBX194LA,
             LBX196LA, LBX206LA, LBX209LA,

             LBXBCD, LBXBPB, LBXTHG,

             DRTACAR, DRTALCO, DRTATOA, DRTATOC, DRTB12A, DRTBCAR, DRTCAFF,
DRTCALC, DRTCARB,
             DRTCOPP, DRTCryp, DRTFA, DRTFDfE, DRTFF, DRTFIBE, DRTFOLA,
DRTIRON,
             DRTLCO, DRTLZ, DRTM161, DRTM181, DRTM201, DRTM221, DRTMAGN,
DRTMFAT, DRTNIAC,
             DRTP182, DRTP183, DRTP184, DRTP204, DRTP205, DRTP225, DRTP226,
DRTPFAT, DRTPHOS,
             DRTPOTA, DRTPROT, DRTRET, DRTS040, DRTS060, DRTS080, DRTS100,
DRTS120, DRTS140,
             DRTS160, DRTS180, DRTSELE, DRTSFAT, DRTSODI, DRTSUGR, DRTTFAT,
DRTTHEO, DRTVARA,
             DRTVB1, DRTVB2, DRTVB6, DRTVB12, DRTVC, DRTVK, DRTZINC);
run;

```

```
data app.pcb_nutr_data_nomiss;  
set app.pcb_nutr_data;  
where nmiss=0;  
run;
```

```

/***/**/**/**/**/**/**/**/**QUARTILE SCORE VARIABLES**/**/**/**/**/**/**/**/**/**/**/;
proc rank data=app.pcb_nutr_data_nomiss group=4 out=pcb_nut_ranked;
var
    LBX028LA LBX066LA LBX074LA LBX105LA LBX118LA LBX156LA
    LBXD03LA LBXD05LA LBXD07LA LBXF08LA LBXPCBLA
    LBX044LA LBX049LA LBX052LA LBX087LA LBX099LA LBX101LA LBX110LA
LBX138LA
    LBX146LA LBX149LA LBX151LA LBX153LA LBX170LA LBX177LA LBX178LA
LBX180LA LBX183LA LBX187LA LBX194LA
    LBX196LA LBX206LA LBX209LA


    LBXBCD LBXBPB LBXTHG


    DRTACAR DRTALCO DRTATOA DRTATOC DRTB12A DRTBCAR DRTCFAFF
DRTCALC DRTCARB
    DRTCOPP DRTCRYP DRTFA DRTFD FE DRTFF DRTFIBE DRTFOLA DRTIRON
DRTKCAL
    DRTLCO DRTLZ DRTM161 DRTM181 DRTM201 DRTM221 DRTMAGN
DRTMFAT DRTNIAC
    DRTP182 DRTP183 DRTP184 DRTP204 DRTP205 DRTP225 DRTP226 DRTPFAT
DRTPHOS
    DRTPOTA DRTPROT DRTRET DRTS040 DRTS060 DRTS080 DRTS100 DRTS120
DRTS140
    DRTS160 DRTS180 DRTSELE DRTSFAT DRTSODI DRTSUGR DRTTFAT
DRTTHEO DRTVARA
    DRTVB1 DRTVB2 DRTVB6 DRTVB12 DRTVC DRTVK DRTZINC
;
Ranks
    LBX028LAq LBX066LAq LBX074LAq LBX105LAq LBX118LAq LBX156LAq
    LBXD03LAq LBXD05LAq LBXD07LAq LBXF08LAq LBXPCBLaq
    LBX044LAq LBX049LAq LBX052LAq LBX087LAq LBX099LAq LBX101LAq
LBX110LAq LBX138LAq
    LBX146LAq LBX149LAq LBX151LAq LBX153LAq LBX170LAq LBX177LAq
LBX178LAq LBX180LAq LBX183LAq LBX187LAq LBX194LAq
    LBX196LAq LBX206LAq LBX209LAq

```

LBXBCDq LBXBPBq LBXTHGq

DRTACARq DRTALCOq DRTATOAq DRTATOCq DRTB12Aq DRTBCARq
DRTCAFFq DRTCALCq DRTCARBq
DRTCOPPq DRTCrypPq DRTFAq DRTFDfEq DRTFFq DRTFIBEq DRTFOLAq
DRTIRONq DRTKCALq
DRTLCOq DRTLZq DRTM161q DRTM181q DRTM201q DRTM221q DRTMAGNq
DRTMFATq DRTNIACq
DRTp182q DRTp183q DRTp184q DRTp204q DRTp205q DRTp225q DRTp226q
DRTPFATq DRTPHOSq
DRTPOTAq DRTPROTq DRTRETq DRTS040q DRTS060q DRTS080q DRTS100q
DRTS120q DRTS140q
DRTS160q DRTS180q DRTSELEq DRTSFATq DRTSODIq DRTSUGRq DRTTFATq
DRTTHEOq DRTVARAq
DRTVB1q DRTVB2q DRTVB6q DRTVB12q DRTVCq DRTVKq DRTZINCq;
run;

proc freq data=pcb_nut_ranked;
table LBX028LAq LBX066LAq LBX074LAq LBX105LAq LBX118LAq LBX156LAq
LBXD03LAq LBXD05LAq LBXD07LAq LBXF08LAq LBXPCBLAq
LBX044LAq LBX049LAq LBX052LAq LBX087LAq LBX099LAq LBX101LAq
LBX110LAq LBX138LAq
LBX146LAq LBX149LAq LBX151LAq LBX153LAq LBX170LAq LBX177LAq
LBX178LAq LBX180LAq
LBX183LAq LBX187LAq LBX194LAq LBX196LAq LBX206LAq LBX209LAq

LBXBCDq LBXBPBq LBXTHGq

DRTACARq DRTALCOq DRTATOAq DRTATOCq DRTB12Aq DRTBCARq
DRTCAFFq DRTCALCq DRTCARBq
DRTCOPPq DRTCrypPq DRTFAq DRTFDfEq DRTFFq DRTFIBEq DRTFOLAq
DRTIRONq DRTKCALq
DRTLCOq DRTLZq DRTM161q DRTM181q DRTM201q DRTM221q DRTMAGNq
DRTMFATq DRTNIACq
DRTp182q DRTp183q DRTp184q DRTp204q DRTp205q DRTp225q DRTp226q
DRTPFATq DRTPHOSq
DRTPOTAq DRTPROTq DRTRETq DRTS040q DRTS060q DRTS080q DRTS100q
DRTS120q DRTS140q
DRTS160q DRTS180q DRTSELEq DRTSFATq DRTSODIq DRTSUGRq DRTTFATq
DRTTHEOq DRTVARAq
DRTVB1q DRTVB2q DRTVB6q DRTVB12q DRTVCq DRTVKq DRTZINCq;
run;

/*/*/*/*/*/*Check distribution of ALT and Log Transform if Needed/*/*/*/*/*/*/*/*;

```

proc univariate data=pcb_nut_ranked noprint;
  histogram lbxsatsi;
run;

```

```

data pcb_nut_ranked;
  set pcb_nut_ranked;
  logALT=log(lbxsatsi);
run;

```

```

proc univariate data=pcb_nut_ranked;
  histogram logALT;
run;

```

```

/*/*/*/*/*/*Check Correlation of Ranked Variables *//*/*/*/*/*/*;

```

```

proc corr data=pcb_nut_ranked spearman;
  var logalt LBX028LAq LBX066LAq LBX074LAq LBX105LAq LBX118LAq LBX156LAq
    LBXD03LAq LBXD05LAq LBXD07LAq LBXF08LAq LBXPCBLAq
    LBX044LAq LBX049LAq LBX052LAq LBX087LAq LBX099LAq LBX101LAq
    LBX110LAq LBX138LAq
    LBX146LAq LBX149LAq LBX151LAq LBX153LAq LBX170LAq LBX177LAq
    LBX178LAq LBX180LAq
    LBX183LAq LBX187LAq LBX194LAq LBX196LAq LBX206LAq LBX209LAq
    LBXBCDq LBXBPBq LBXTHGq;
run;

```

```

proc corr data=pcb_nut_ranked spearman;
  var logalt LBXBCD LBXBPB LBXTHG ;
run;

```

```

/*NO ALCO, ATOA, B12A, TP184*/

```

```

proc corr data=pcb_nut_ranked spearman;
  var logalt
    DRTACARq DRTATOCq DRTBCARq DRTCAFFq DRTCALCq DRTCARBq
    DRTCOPPq DRTCrypPq DRTFAq DRTFDfEq DRTFFq DRTFIBEq DRTFOLAq
    DRTIRONq
    DRTLCOq DRTLZq DRTM161q DRTM181q DRTM201q DRTM221q DRTMAGNq
    DRTMFATq DRTNIACq
    DRTP182q DRTP183q DRTP204q DRTP205q DRTP225q DRTP226q DRTPFATq
    DRTPHOSq
    DRTPOTAq DRTPROTq DRTRETq DRTS040q DRTS060q DRTS080q DRTS100q
    DRTS120q DRTS140q
    DRTS160q DRTS180q DRTSELEq DRTSFATq DRTSODIq DRTSUGRq DRTTFATq
    DRTTHEOq DRTVARAq
    DRTVB1q DRTVB2q DRTVB6q DRTVB12q DRTVCq DRTVKq DRTZINCq;
run;

```

```

/*/*/*/*/*/*SPLIT INTO TEST/VALIDATE DATASETS*//*/*/*/*/*/*;

```

```

DATA pcb_nut_ranked;
set pcb_nut_ranked;
i=ranuni(99);
run;
proc sort data=pcb_nut_ranked; by i; run;
data pcb_nut_ranked;
set pcb_nut_ranked;
obs=_n_;
run;

```

```

data pcb_nut_ranked;
set pcb_nut_ranked;
if obs<=(928/2) then sample=1;
if obs>(928/2) then sample=0;
if ridreth1=3 then bin_race=0;
else bin_race=1;
run;
run;

```

```

/*Test Covariates Across 2 Datasets*/
%macro test(var);
Proc genmod data=pcb_nut_ranked;
class sample;
model sample=&var /type3;
run;
%mend;

```

```

%test(bin_race);
%test(RIDAGEYR);
%test(riagendr);
%test(indfmpir);
%test(bmxbmi);
%test(logalt);

```

```

proc freq data=pcb_nut_ranked;
table riagendr*sample;
table bin_race*sample;
run;

```

```

/*Create TEST Dataset*/
data app.test;
set pcb_nut_ranked;
where sample=1;
run;

```

```

/*Create VALIDATION Dataset*/

```



```

data app.validate;
set pcb_nut_ranked;
where sample=0;
run;

```

```

data app.test;
set app.test;
if ridreth1=3 then bin_race=0;
else bin_race=1;
run;

```

```

data app.validate;
set app.validate;
if ridreth1=3 then bin_race=0;
else bin_race=1;
run;

```

```

proc freq data=app.validate;
table bin_race;
run;

```

```

data app.test_small;
set app.test;
keep

```

```

/*outcome*/
logALT seqn

```

```

/*covariates*/
bin_race ridreth1 RIDAGEYR riagendr indfmpir bmx bmi

```

```

/*nutrients quartile score*/
DRTACARq DRTATOCq DRTBCARq DRTCAFFq DRTCALCq DRTCARBq
DRTCOPPq DRTCRYPq DRTFAq DRTFDFFq DRTFFq DRTFIBEq DRTFOLAq
DRTIRONq
DRTLCOq DRTLZq DRTM161q DRTM181q DRTM201q DRTM221q DRTMAGNq
DRTMFATq DRTNIACq
DRTPI82q DRTPI83q DRTPI204q DRTPI205q DRTPI225q DRTPI226q DRTPIFATq
DRTPHOSq
DRTPIOTAq DRTPIROTq DRTRETq DRTS040q DRTS060q DRTS080q DRTS100q
DRTS120q DRTS140q
DRTS160q DRTS180q DRTSELEq DRTSFATq DRTSODIq DRTSUGRq DRTTFATq
DRTTHEOq DRTVARAq
DRTVB1q DRTVB2q DRTVB6q DRTVB12q DRTVCq DRTVKq DRTZINCq

```

```

/*chems quartile score*/
LBX028LAq LBX066LAq LBX074LAq LBX105LAq LBX118LAq LBX156LAq
LBXD03LAq LBXD05LAq LBXD07LAq LBXF08LAq LBXPCBLAq

```

```

    LBX044LAq LBX049LAq LBX052LAq LBX087LAq LBX099LAq LBX101LAq
    LBX110LAq LBX138LAq
    LBX146LAq LBX149LAq LBX151LAq LBX153LAq LBX170LAq LBX177LAq
    LBX178LAq LBX180LAq
    LBX183LAq LBX187LAq LBX194LAq LBX196LAq LBX206LAq LBX209LAq

    LBXBCDq LBXBPBq LBXTHGq

;

run;

data app.validate_small;
set app.validate;
keep

/*outcome*/
logALT seqn

/*covariates*/
bin_race ridreth1 RIDAGEYR riagendr indfmpir bmx bmi

/*nutrients quartile score*/
    DRTACARq DRTATOCq DRTBCARq DRTCAFFq DRTCALCq DRTCARBq
    DRTCOPPq DRTCRYPq DRTFAq DRTFDFFeq DRTFFq DRTFIBEq DRTFOLAq
    DRTIRONq
    DRTLYCOq DRTLZq DRTM161q DRTM181q DRTM201q DRTM221q DRTMAGNq
    DRTMFATq DRTNIACq
    DRTP182q DRTP183q DRTP204q DRTP205q DRTP225q DRTP226q DRTPFATq
    DRTPHOSq
    DRTPOTAq DRTPROTq DRTRETq DRTS040q DRTS060q DRTS080q DRTS100q
    DRTS120q DRTS140q
    DRTS160q DRTS180q DRTSELEq DRTSFATq DRTSODIq DRTSUGRq DRTTFATq
    DRTTHEOq DRTVARAq
    DRTVB1q DRTVB2q DRTVB6q DRTVB12q DRTVCq DRTVKq DRTZINCq

/*chems quartile score*/
    LBX028LAq LBX066LAq LBX074LAq LBX105LAq LBX118LAq LBX156LAq
    LBXD03LAq LBXD05LAq LBXD07LAq LBXF08LAq LBXPCBLAq

    LBX044LAq LBX049LAq LBX052LAq LBX087LAq LBX099LAq LBX101LAq
    LBX110LAq LBX138LAq
    LBX146LAq LBX149LAq LBX151LAq LBX153LAq LBX170LAq LBX177LAq
    LBX178LAq LBX180LAq
    LBX183LAq LBX187LAq LBX194LAq LBX196LAq LBX206LAq LBX209LAq

```

```

        LBXBCDq LBXBPBq LBXTHGq

;

run;

/**/**/**/**/**/**ESTIMATE Chems WEIGHTS FROM TEST DATASET**/**/**/**/**/**/;
/* b1*RIDRETH1 - b2*RIDAGEYR - b3*riagendr- -b4*indfmpir-b5*bmxbmi*/
data start;
    _type_='PARMS';
    alpha=2; Chems=0.10001; b1=0.1; b2=-0.1; b3=-0.2; b4=0.1; b5=0.01; sigma=1.0;
    array inwts

        WLBX028LAq WLBX066LAq WLBX074LAq WLBX105LAq WLBX118LAq
        WLBX156LAq
        WLBXD03LAq WLBXD05LAq WLBXD07LAq WL BXF08LAq WL BXPCBLAq
        WL BX044LAq WL BX049LAq WL BX052LAq WL BX087LAq WL BX099LAq
        WL BX101LAq WL BX110LAq WL BX138LAq
        WL BX146LAq WL BX149LAq WL BX151LAq WL BX153LAq WL BX170LAq
        WL BX177LAq WL BX178LAq WL BX180LAq
        WL BX183LAq WL BX187LAq WL BX194LAq WL BX196LAq WL BX206LAq
        WL BX209LAq

        WLXBCDq WLXBPBq WLXTHGq

;

    do over inwts;
        inwts=1/36;
        end;

proc nlp data=app.test_small technique=trureg
    maxiter=10000 maxfunc=10000 inest=start outest=outstuff nomiss;
*objective function;
    max logL;
*define parameters; *nutr;
    parms alpha CHEMS b1 b2 b3 b4 b5

        WL BX028LAq WL BX066LAq WL BX074LAq WL BX105LAq WL BX118LAq
        WL BX156LAq
        WL BXD03LAq WL BXD05LAq WL BXD07LAq WL BXF08LAq WL BXPCBLAq
        WL BX044LAq WL BX049LAq WL BX052LAq WL BX087LAq WL BX099LAq
        WL BX101LAq WL BX110LAq WL BX138LAq

```

WLBX146LAq WLBX149LAq WLBX151LAq WLBX153LAq WLBX170LAq
 WLBX177LAq WLBX178LAq WLBX180LAq
 WLBX183LAq WLBX187LAq WLBX194LAq WLBX196LAq WLBX206LAq
 WLBX209LAq
 WLBXBCDq WLBXBPBq WLBXTHGq
 ;

*program statements;

logL= -0.5*log(sigma)-0.5*(1/sigma)*(logALT-alpha-
 - b1*bin_race - b2*RIDAGEYR - b3*riagendr- -b4*indfmpir-b5*bmxbmi
 -chems * (WLBXBCDq*LBXBCDq+ WLBXBPBq*LBXBPBq+
 WLBXTHGq*LBXTHGq +WLBX028LAq*LBX028LAq+ WLBX066LAq*LBX066LAq+
 WLBX074LAq*LBX074LAq+ WLBX105LAq*LBX105LAq+
 WLBX118LAq*LBX118LAq+
 WLBX156LAq*LBX156LAq+ WLBXD03LAq*LBXD03LAq+
 WLBXD05LAq*LBXD05LAq+
 WLBXD07LAq*LBXD07LAq+
 WLBXF08LAq*LBXF08LAq+ WLBXPCBLAq*LBXPCBLAq+
 WLBX044LAq*LBX044LAq+
 WLBX049LAq*LBX049LAq+
 WLBX052LAq*LBX052LAq+ WLBX087LAq*LBX087LAq+ WLBX099LAq*LBX099LAq+
 WLBX101LAq*LBX101LAq+
 WLBX110LAq*LBX110LAq+ WLBX138LAq*LBX138LAq+ WLBX146LAq*LBX146LAq+
 WLBX149LAq*LBX149LAq+
 WLBX151LAq*LBX151LAq+ WLBX153LAq*LBX153LAq+ WLBX170LAq*LBX170LAq+
 WLBX177LAq*LBX177LAq+
 WLBX178LAq*LBX178LAq+ WLBX180LAq*LBX180LAq+ WLBX183LAq*LBX183LAq+
 WLBX187LAq*LBX187LAq+
 WLBX194LAq*LBX194LAq+ WLBX196LAq*LBX196LAq+ WLBX206LAq*LBX206LAq+
 WLBX209LAq*LBX209LAq))**2;

*linear constraints;

lincon

/*Weights for Chems sum to 1*/

WLBX028LAq+ WLBX066LAq+ WLBX074LAq+
 WLBX105LAq+
 WLBX118LAq+ WLBX156LAq+ WLBXD03LAq+
 WLBXD05LAq+
 WLBXD07LAq+ WLBXF08LAq+ WLBXPCBLAq+
 WLBX044LAq+
 WLBX049LAq+ WLBX052LAq+ WLBX087LAq+
 WLBX099LAq+
 WLBX101LAq+ WLBX110LAq+ WLBX138LAq+
 WLBX146LAq+

```

WLBX170LAq+
WLBX183LAq+
WLBX206LAq+
WLBXTHGq= 1;
WLBX149LAq+ WLBX151LAq+ WLBX153LAq+
WLBX177LAq+ WLBX178LAq+ WLBX180LAq+
WLBX187LAq+ WLBX194LAq+ WLBX196LAq+
WLBX209LAq+ WLBXBCDq+ WLBXBPBq+

```

```

*bounds on weights: all in (0,1);
  bounds

```

```

0<WLBX028LAq<1, 0<WLBX066LAq<1,
0<WLBX074LAq<1, 0<WLBX105LAq<1,
0<WLBX118LAq<1, 0<WLBX156LAq<1,
0<WLBXD03LAq<1, 0<WLBXD05LAq<1,
0<WLBXD07LAq<1, 0<WLBXF08LAq<1,
0<WLBXPCBLAq<1, 0<WLBX044LAq<1,
0<WLBX049LAq<1, 0<WLBX052LAq<1,
0<WLBX087LAq<1, 0<WLBX099LAq<1,
0<WLBX101LAq<1, 0<WLBX110LAq<1,
0<WLBX138LAq<1, 0<WLBX146LAq<1,
0<WLBX149LAq<1, 0<WLBX151LAq<1,
0<WLBX153LAq<1, 0<WLBX170LAq<1,
0<WLBX177LAq<1, 0<WLBX178LAq<1,
0<WLBX180LAq<1, 0<WLBX183LAq<1,
0<WLBX187LAq<1, 0<WLBX194LAq<1,
0<WLBX196LAq<1, 0<WLBX206LAq<1,
0<WLBX209LAq<1, 0<WLBXBCDq<1,
0<WLBXBPBq<1, 0<WLBXTHGq<1,

```

```

/*determine set of "protective" nutrients by constraining vits to be negative and "negative" chems
by constraining pcbs positive*/

```

```

  chems>0.01;

```

```

run;

```

```

/***/**/**/**/**/**VALIDATE WEIGHTS IN VALIDATE DATASET**/**/**/**/**/**/**/**/;
/*Populate dataset with weights and set MERGE-BY Variable*/

```

```

data outstuff;
set work.outstuff;
where _type_='PARMS';
sample=1;
run;

```

```

data validate;

```



```

reps=1000 seed=113084 outhits out=work.boot_test;
run;

```

```

data start;
  _type_='PARMS';
  alpha=2; Chems=0.10001; b1=0.1; b2=0.1; b3=-0.2; b4=0.1; b5=0.01; b6=-0.1; sigma=1.0;
  array inwts

```

```

    WLBX028LAq WLBX066LAq WLBX074LAq WLBX105LAq WLBX118LAq
WLBX156LAq
    WLBXD03LAq WLBXD05LAq WLBXD07LAq WLBXF08LAq WLBXPCBLAq
    WLBX044LAq WLBX049LAq WLBX052LAq WLBX087LAq WLBX099LAq
WLBX101LAq WLBX110LAq WLBX138LAq
    WLBX146LAq WLBX149LAq WLBX151LAq WLBX153LAq WLBX170LAq
WLBX177LAq WLBX178LAq WLBX180LAq
    WLBX183LAq WLBX187LAq WLBX194LAq WLBX196LAq WLBX206LAq
WLBX209LAq WLBXBCDq WLBXBPBq WLBXTHGq ;

```

```

    do over inwts;
      inwts=1/36;
    end;

```

```

proc nlp data=boot_test technique=trureg
  maxiter=10000 maxfunc=10000 inest=start outest=outstuff nomiss ;
  *set by variable for bootstrap samples;
  by replicate;
  *objective function;
    max logL;
  *define parameters;
    parms alpha chems b1 b2 b3 b4 b5 b6
    WLBX028LAq WLBX066LAq WLBX074LAq WLBX105LAq WLBX118LAq
WLBX156LAq
    WLBXD03LAq WLBXD05LAq WLBXD07LAq WLBXF08LAq WLBXPCBLAq
    WLBX044LAq WLBX049LAq WLBX052LAq WLBX087LAq WLBX099LAq
WLBX101LAq WLBX110LAq WLBX138LAq
    WLBX146LAq WLBX149LAq WLBX151LAq WLBX153LAq WLBX170LAq
WLBX177LAq WLBX178LAq WLBX180LAq
    WLBX183LAq WLBX187LAq WLBX194LAq WLBX196LAq WLBX206LAq
WLBX209LAq WLBXBCDq WLBXBPBq WLBXTHGq;

  *program statements;
    logL= -0.5*log(sigma)-0.5*(1/sigma)*(logALT-alpha-
    - b1*bin_race - b2*RIDAGEYR-b6*ridageyr*ridageyr - b3*riagendr- -b4*indfmpir-
b5*bmxbmi

```

```

-chems * ( WLBX028LAq*LBX028LAq+
WLBX066LAq*LBX066LAq+ WLBX074LAq*LBX074LAq+ WLBX105LAq*LBX105LAq+
WLBX118LAq*LBX118LAq+
WLBX156LAq*LBX156LAq+ WLBXD03LAq*LBXD03LAq+
WLBXD05LAq*LBXD05LAq+
WLBXD07LAq*LBXD07LAq+
WLBXF08LAq*LBXF08LAq+ WLBXPCBLAq*LBXPCBLAq+
WLBX044LAq*LBX044LAq+
WLBX049LAq*LBX049LAq+
WLBX052LAq*LBX052LAq+ WLBX087LAq*LBX087LAq+ WLBX099LAq*LBX099LAq+
WLBX101LAq*LBX101LAq+
WLBX110LAq*LBX110LAq+ WLBX138LAq*LBX138LAq+ WLBX146LAq*LBX146LAq+
WLBX149LAq*LBX149LAq+
WLBX151LAq*LBX151LAq+ WLBX153LAq*LBX153LAq+ WLBX170LAq*LBX170LAq+
WLBX177LAq*LBX177LAq+
WLBX178LAq*LBX178LAq+ WLBX180LAq*LBX180LAq+ WLBX183LAq*LBX183LAq+
WLBX187LAq*LBX187LAq+
WLBX194LAq*LBX194LAq+ WLBX196LAq*LBX196LAq+ WLBX206LAq*LBX206LAq+
WLBX209LAq*LBX209LAq+
WLBXBCDq*LBXBCDq+ WLBXBPBq*LBXBPBq+ WLBXTHGq*LBXTHGq))**2;

```

*linear constraints;

lincon

/*Weights for chems sum to 1;*/

```

WLBX028LAq+ WLBX066LAq+ WLBX074LAq+
WLBX105LAq+
WLBX118LAq+ WLBX156LAq+ WLBXD03LAq+
WLBXD05LAq+
WLBXD07LAq+ WLBXF08LAq+ WLBXPCBLAq+
WLBX044LAq+
WLBX049LAq+ WLBX052LAq+ WLBX087LAq+
WLBX099LAq+
WLBX101LAq+ WLBX110LAq+ WLBX138LAq+
WLBX146LAq+
WLBX149LAq+ WLBX151LAq+ WLBX153LAq+
WLBX170LAq+
WLBX177LAq+ WLBX178LAq+ WLBX180LAq+
WLBX183LAq+
WLBX187LAq+ WLBX194LAq+ WLBX196LAq+
WLBX206LAq+
WLBX209LAq+ WLBXBCDq+ WLBXBPBq +
WLBXTHGq= 1; */

```

*bounds on weights: all in (0,1);

bounds


```

0<WLBX028LAq<1, 0<WLBX066LAq<1,
0<WLBX074LAq<1, 0<WLBX105LAq<1,
0<WLBX118LAq<1, 0<WLBX156LAq<1,
0<WLBXD03LAq<1, 0<WLBXD05LAq<1,
0<WLBXD07LAq<1, 0<WLBXF08LAq<1,
0<WLBXPCBLAq<1, 0<WLBX044LAq<1,
0<WLBX049LAq<1, 0<WLBX052LAq<1,
0<WLBX087LAq<1, 0<WLBX099LAq<1,
0<WLBX101LAq<1, 0<WLBX110LAq<1,
0<WLBX138LAq<1, 0<WLBX146LAq<1,
0<WLBX149LAq<1, 0<WLBX151LAq<1,
0<WLBX153LAq<1, 0<WLBX170LAq<1,
0<WLBX177LAq<1, 0<WLBX178LAq<1,
0<WLBX180LAq<1, 0<WLBX183LAq<1,
0<WLBX187LAq<1, 0<WLBX194LAq<1,
0<WLBX196LAq<1, 0<WLBX206LAq<1,
0<WLBX209LAq<1, 0<WLBXBCDq<1,
0<WLBXBPBq<1, 0<WLBXTHGq<1,

```

```

/*determine set of "protective" nutrients by constraining vits to be negative and "negative" chems
by constraining pcbs positive*/

```

```

chems>0.01;

```

```

run;

```

```

/***/**/**/**/**/**VALIDATE WEIGHTS IN VALIDATE DATASET**/**/**/**/**/**/**/;
/*Populate dataset with weights and set MERGE-BY Variable*/

```

```

/*Keep only Parameter Estimates*/

```

```

data work.outstuff;
set work.outstuff;
where _type_='PARMS';
run;

```

```

/*Validate bootstrap samples in Validate dataset*/
/*

```

```

data replicates;
do replicate=1 to 1000;
do obs=1 to 464; output;
end;
end;
run;

```

```

data valid_rep;
set app.validate_small;
obs=_n_;

```

```
run;
proc sort data=work.valid_rep;
by obs;
run;
```

```
proc sort data=work.replicates;
by obs;
run;
```

```
data valid_rep_tog;
merge valid_rep replicates;
by obs;
run;
```

```
proc sort data=valid_rep_tog;;
by replicate;
run;
```

```
proc sort data=outstuff;
by replicate;
run;
*/
```

```
/*Validate bootstrap samples in bootstrap samples dataset*/
```

```
/*VALIDATE WEIGHTS FROM TEST DATASET*/
```

```
data together;
    merge boot_test outstuff; by replicate;
        chems =      WLBX028LAq*LBX028LAq+ WLBX066LAq*LBX066LAq+
WLBX074LAq*LBX074LAq+ WLBX105LAq*LBX105LAq+
                        WLBX118LAq*LBX118LAq+
WLBX156LAq*LBX156LAq+  WLBXD03LAq*LBXD03LAq+
WLBXD05LAq*LBXD05LAq+
                        WLBXD07LAq*LBXD07LAq+
WLBXF08LAq*LBXF08LAq+ WLBXPCBLAq*LBXPCBLAq+
WLBX044LAq*LBX044LAq+
                        WLBX049LAq*LBX049LAq+
WLBX052LAq*LBX052LAq+ WLBX087LAq*LBX087LAq+ WLBX099LAq*LBX099LAq+
                        WLBX101LAq*LBX101LAq+
WLBX110LAq*LBX110LAq+ WLBX138LAq*LBX138LAq+ WLBX146LAq*LBX146LAq+
                        WLBX149LAq*LBX149LAq+
WLBX151LAq*LBX151LAq+ WLBX153LAq*LBX153LAq+ WLBX170LAq*LBX170LAq+
                        WLBX177LAq*LBX177LAq+
WLBX178LAq*LBX178LAq+ WLBX180LAq*LBX180LAq+ WLBX183LAq*LBX183LAq+
                        WLBX187LAq*LBX187LAq+
WLBX194LAq*LBX194LAq+ WLBX196LAq*LBX196LAq+ WLBX206LAq*LBX206LAq+
```

```

                                WLBX209LAq*LBX209LAq+
WLBXBCDq*LBXBCDq+ WLBXBPBq*LBXBPBq+ WLBXTHGq*LBXTHGq;

```

```

run;

```

```

proc genmod data=together ;
by replicate;
class bin_race riagendr ;
  model logalt = chems bin_race ridageyr ridageyr*ridageyr riagendr indfmpir
bmxbmi/type3;/*nutr*/
  ods output ParameterEstimates=parms;
run;

```

```

data chems;
set work.parms;
where parameter='chems';
if probchisq<=0.05 then power=1;
else power=0;
run;

```

```

proc freq data=chems;
tables power;
run;

```

```

data chems_weights;
merge chems outstuff;
by replicate;
run;

```

```

ods rtf file="C:\Users\Caroline\Documents\Dissertation\WORK\means_chem.rtf";
proc means data=chems_weights;
where power=1;
var WLBX028LAq WLBX066LAq WLBX074LAq WLBX105LAq WLBX118LAq
WLBX156LAq
    WLBXD03LAq WLBXD05LAq WLBXD07LAq WLBXF08LAq WLBXPCBLAq
    WLBX044LAq WLBX049LAq WLBX052LAq WLBX087LAq WLBX099LAq
WLBX101LAq WLBX110LAq WLBX138LAq
    WLBX146LAq WLBX149LAq WLBX151LAq WLBX153LAq WLBX170LAq
WLBX177LAq WLBX178LAq WLBX180LAq
    WLBX183LAq WLBX187LAq WLBX194LAq WLBX196LAq WLBX206LAq
WLBX209LAq WLBXBCDq WLBXBPBq WLBXTHGq;
  ods output summary=means;

run;
ods rtf close;

```

```

data validate;
set app.validate_small;
dummy=1;
run;

data means;
set work.means;
dummy=1;
run;

data val_means;
merge validate means;
by dummy;
run;
data val_means;
set val_means;
chems_mean=
          WLBX028LAq_mean*LBX028LAq+
WLBX066LAq_mean*LBX066LAq+ WLBX074LAq_mean*LBX074LAq+
WLBX105LAq_mean*LBX105LAq+
          WLBX118LAq_mean*LBX118LAq+
WLBX156LAq_mean*LBX156LAq+ WLBXD03LAq_mean*LBXD03LAq+
WLBXD05LAq_mean*LBXD05LAq+
          WLBXD07LAq_mean*LBXD07LAq+
WLBXF08LAq_mean*LBXF08LAq+ WLBXPCBLAq_mean*LBXPCBLAq+
WLBX044LAq_mean*LBX044LAq+
          WLBX049LAq_mean*LBX049LAq+
WLBX052LAq_mean*LBX052LAq+ WLBX087LAq_mean*LBX087LAq+
WLBX099LAq_mean*LBX099LAq+
          WLBX101LAq_mean*LBX101LAq+
WLBX110LAq_mean*LBX110LAq+ WLBX138LAq_mean*LBX138LAq+
WLBX146LAq_mean*LBX146LAq+
          WLBX149LAq_mean*LBX149LAq+
WLBX151LAq_mean*LBX151LAq+ WLBX153LAq_mean*LBX153LAq+
WLBX170LAq_mean*LBX170LAq+
          WLBX177LAq_mean*LBX177LAq+
WLBX178LAq_mean*LBX178LAq+ WLBX180LAq_mean*LBX180LAq+
WLBX183LAq_mean*LBX183LAq+
          WLBX187LAq_mean*LBX187LAq+
WLBX194LAq_mean*LBX194LAq+ WLBX196LAq_mean*LBX196LAq+
WLBX206LAq_mean*LBX206LAq+
          WLBX209LAq_mean*LBX209LAq+
WLBXBCDq_mean*LBXBCDq+   WLBXBPBq_mean*LBXBPBq+
WLBXTHGq_mean*LBXTHGq;
;

run;

proc genmod data=val_means;

```

```

class bin_race riagendr ;
  model logalt = chems_mean bin_race RIDAGEYR ridageyr*ridageyr riagendr indfmpir
  bmx bmi/type3;
  ods output ParameterEstimates=parms;
  run;

```

```

  data app.chems_finalmodel;
  set work.parms;
  run;

```

```

proc univariate data=chems_weights noprint;
  where power=1;
  histogram WLBX028LAq WLBX066LAq WLBX074LAq WLBX105LAq WLBX118LAq
  WLBX156LAq
  WLBXD03LAq WLBXD05LAq WLBXD07LAq WLBXF08LAq WLBXPCBLAq
  WLBX044LAq WLBX049LAq WLBX052LAq WLBX087LAq WLBX099LAq
  WLBX101LAq WLBX110LAq WLBX138LAq
  WLBX146LAq WLBX149LAq WLBX151LAq WLBX153LAq WLBX170LAq
  WLBX177LAq WLBX178LAq WLBX180LAq
  WLBX183LAq WLBX187LAq WLBX194LAq WLBX196LAq WLBX206LAq
  WLBX209LAq WLBXBCDq WLBXBPBq WLBXTHGq;
  inset n mean p5='5th Percentile' p95='95th Percentile' / pos = ne height=4.0
  format=best4.;
  run;

```

```

/*/*/*/*/*/*/*/*ESTIMATE NUTR WEIGHTS FROM TEST DATASET/*/*/*/*/*/*/*/*;

```

```

data start;
  _type_='PARMS';
  alpha=2; nutr=0.10001; b1=0.1; b2=0.1; b3=-0.2; b4=0.1; b5=0.01; b6=-.1 ; sigma=1.0;
  array inwts

```

```

  WDR TACARq WDR TATOCq WDR TBCARq WDR TCAFFq WDR TCA LCq
  WDR TCARBq
  WDR TCOPPq WDR TCRYPq WDR TFAq WDR TFD FEq WDR TFFq WDR TFIBEq
  WDR TFOLAq WDR TIRONq
  WDR TLYCOq WDR T LZq WDR TM161q WDR TM181q WDR TM201q WDR TM221q
  WDR TMAGNq WDR TMFATq WDR TN IACq
  WDR TP182q WDR TP183q WDR TP204q WDR TP205q WDR TP225q WDR TP226q
  WDR TP FATq WDR TP HOSq
  WDR TPOTAq WDR TPROTq WDR TRETq WDR TS040q WDR TS060q WDR TS080q
  WDR TS100q WDR TS120q WDR TS140q
  WDR TS160q WDR TS180q WDR TSELEq WDR TSFATq WDR TSODIq
  WDR TSUGRq WDR TT FATq WDR TT HEOq WDR TVARAq
  WDR TVB1q WDR TVB2q WDR TVB6q WDR TVB12q WDR TVCq WDR TVKq
  WDR TZINCq

```

;

```
do over inwts;  
  inwts=1/56;  
end;
```

```
proc nlp data=app.test_small technique=trureg  
  maxiter=10000 maxfunc=10000 inest=start outest=outstuff nomiss;  
*objective function;  
  max logL;  
*define parameters; *nutr;  
  parms alpha nutr b1 b2 b3 b4 b5  
    WDRTACARq WDRATATOCq WDRTBCARq WDRTCAFFq WDRTCALCq  
WDRTCARBq  
    WDRTCOPPq WDRTCRYPq WDRTFAq WDRTFDFEq WDRTFFq WDRTFIBEq  
WDRTFOLAq WDRTIRONq  
    WDRTLYCOq WDRTLZq WDRTM161q WDRTM181q WDRTM201q WDRTM221q  
WDRTMAGNq WDRTMFATq WDRTNIAcq  
    WDRTP182q WDRTP183q WDRTP204q WDRTP205q WDRTP225q WDRTP226q  
WDRTPFATq WDRTPHOSq  
    WDRTPOTAq WDRTPROTq WDRTRETq WDRTS040q WDRTS060q WDRTS080q  
WDRTS100q WDRTS120q WDRTS140q  
    WDRTS160q WDRTS180q WDRTSELEq WDRTSFATq WDRTSODIq  
WDRTSUGRq WDRTTFATq WDRTTHEOq WDRTVARAq  
    WDRTVB1q WDRTVB2q WDRTVB6q WDRTVB12q WDRTVCq WDRTVKq  
WDRTZINCq ;  
  
*program statements;  
  logL= -0.5*log(sigma)-0.5*(1/sigma)*(logALT-alpha-  
    - b1*bin_race - b2*RIDAGEYR - b3*riagendr- -b4*indfmpir-b5*bmxbmi-  
b2*RIDAGEYR*RIDAGEYR
```

```
    -nutr* (WDRTACARq*DRTACARq+ WDRATATOCq*DRTATOCq+  
WDRTBCARq*DRTBCARq+ WDRTCAFFq*DRTCAFFq+ WDRTCALCq*DRTCALCq+  
    WDRTCARBq*DRTCARBq+  
    WDRTCOPPq*DRTCOPPq+ WDRTCRYPq*DRTCROPq+ WDRTFAq*DRTFAq+  
WDRTFDFEq*DRTFDFEq+  
    WDRTFFq*DRTFFq+ WDRTFIBEq*DRTFIBEq+  
WDRTFOLAq*DRTFOLAq+ WDRTIRONq*DRTIRONq+ WDRTLYCOq*DRTLYCOq+  
    WDRTLZq*DRTLZq+ WDRTM161q*DRTM161q+  
WDRTM181q*DRTM181q+ WDRTM201q*DRTM201q+ WDRTM221q*DRTM221q+  
    WDRTMAGNq*DRTMAGNq+  
WDRTMFATq*DRTMFATq+ WDRTNIAcq*DRTNIAcq+ WDRTP182q*DRTTP182q+  
WDRTP183q*DRTTP183q+
```

WDRTTP204q*DRTP204q+ WDRTTP205q*DRTP205q+
 WDRTTP225q*DRTP225q+ WDRTTP226q*DRTP226q+ WDRTTPFATq*DRTPFATq+
 WDRTPHOSq*DRTPHOSq+
 WDRTPOTAq*DRTPOTAq+ WDRTPROTq*DRTPROTq+ WDRTRETq*DRTRETq+
 WDRTS040q*DRTS040q+
 WDRTS060q*DRTS060q+ WDRTS080q*DRTS080q+
 WDRTS100q*DRTS100q+ WDRTS120q*DRTS120q+ WDRTS140q*DRTS140q+
 WDRTS160q*DRTS160q+ WDRTS180q*DRTS180q+
 WDRTSELEq*DRTSELEq+ WDRTSFATq*DRTSFATq+ WDRTSODIq*DRTSODIq+
 WDRTSUGRq*DRTSUGRq+
 WDRTTFFATq*DRTTFFATq+ WDRTTHEOq*DRTTHEOq+ WDRTVARAq*DRTVARAg+
 WDRTVB1q*DRTVB1q+
 WDRTVB2q*DRTVB2q+ WDRTVB6q*DRTVB6q+
 WDRTVB12q*DRTVB12q+ WDRTVCq*DRTVCq+ WDRTVKq*DRTVKq+
 WDRTZINCq*DRTZINCq))*2;

*linear constraints;

lincon

/*weights for Nutrients sum to 1*/

WDRTACARq+ WDRTATOCq+ WDRTBCARq+
 WDRTCAFFq+ WDRTCALCq+
 WDRTCARBq+ WDRTCOPPq+ WDRTCRYPq+
 WDRTFAq+ WDRTFDFFeq+
 WDRTFFq+ WDRTFIBEq+ WDRTFOLAq+
 WDRTIRONq+ WDRTLYCOq+
 WDRTLZq+ WDRTM161q+ WDRTM181q+
 WDRTM201q+ WDRTM221q+
 WDRTMAGNq+ WDRTMFATq+ WDRTNIACq+
 WDRTTP182q+ WDRTTP183q+
 WDRTTP204q+ WDRTTP205q+ WDRTTP225q+
 WDRTTP226q+ WDRTTPFATq+
 WDRTPHOSq+ WDRTPOTAq+ WDRTPROTq+
 WDRTRETq+ WDRTS040q+
 WDRTS060q+ WDRTS080q+ WDRTS100q+
 WDRTS120q+ WDRTS140q+
 WDRTS160q+ WDRTS180q+ WDRTSELEq+
 WDRTSFATq+ WDRTSODIq+
 WDRTSUGRq+ WDRTTFFATq+ WDRTTHEOq+
 WDRTVARAq+ WDRTVB1q+
 WDRTVB2q+ WDRTVB6q+ WDRTVB12q+
 WDRTVCq+ WDRTVKq+ WDRTZINCq=1;

*bounds on weights: all in (0,1);

bounds

```

                                0<WDRTACARq<1, 0<WDRTATOCq<1,
0<WDRTBCARq<1, 0<WDRTCAFFq<1, 0<WDRTCALCq<1,
                                0<WDRTCARBq<1, 0<WDRTCOPPq<1,
0<WDRTCRYPq<1, 0<WDRTFAq<1, 0<WDRTFDfEq<1,
                                0<WDRTFFq<1, 0<WDRTFIBEq<1, 0<WDRTFOLAq<1,
0<WDRTIRONq<1, 0<WDRTLCOq<1,
                                0<WDRTLZq<1, 0<WDRTM161q<1, 0<WDRTM181q<1,
0<WDRTM201q<1, 0<WDRTM221q<1,
                                0<WDRTMAGNq<1, 0<WDRTMFATq<1,
0<WDRTNIACq<1, 0<WDRTTP182q<1, 0<WDRTTP183q<1,
                                0<WDRTTP204q<1, 0<WDRTTP205q<1,
0<WDRTTP225q<1, 0<WDRTTP226q<1, 0<WDRTTPFATq<1,
                                0<WDRTPHOSq<1, 0<WDRTPOTAq<1,
0<WDRTPROTq<1, 0<WDRTRETq<1, 0<WDRTS040q<1,
                                0<WDRTS060q<1, 0<WDRTS080q<1,
0<WDRTS100q<1, 0<WDRTS120q<1, 0<WDRTS140q<1,
                                0<WDRTS160q<1, 0<WDRTS180q<1,
0<WDRTSELEq<1, 0<WDRTSFATq<1, 0<WDRTSODIq<1,
                                0<WDRTSUGRq<1, 0<WDRTTFATq<1,
0<WDRTTHEOq<1, 0<WDRTVARAq<1, 0<WDRTVB1q<1,
                                0<WDRTVB2q<1, 0<WDRTVB6q<1,
0<WDRTVB12q<1, 0<WDRTVCq<1, 0<WDRTVKq<1, 0<WDRTZINCq<1,

```

```

/*determine set of "protective" nutrients by constraining vits to be negative and "negative" chems
by constraining pcbs positive*/

```

```

    nutr>0.05;

```

```

run;

```

```

/**/**/**/**/**/**/**VALIDATE WEIGHTS IN VALIDATE DATASET**/**/**/**/**/**/**/**/;
/*Populate dataset with weights and set MERGE-BY Variable*/
/*

```

```

data outstuff;
set work.outstuff;
where _type_='PARMS';
sample=1;
run;

```

```

data validate;
set app.validate;
sample=1;
run;

```



```

data VALIDATE;
set VALIDATE;
dummy=1;
run;

data means;
set test_means;
dummy=1;
run;

data VALIDATE;
merge VALIDATE means;
by dummy;
run;
data VALIDATE;
set VALIDATE;
chems_mean=
          WLBX028LAq_mean*LBX028LAq+
WLBX066LAq_mean*LBX066LAq+ WLBX074LAq_mean*LBX074LAq+
WLBX105LAq_mean*LBX105LAq+
          WLBX118LAq_mean*LBX118LAq+
WLBX156LAq_mean*LBX156LAq+ WLBXD03LAq_mean*LBXD03LAq+
WLBXD05LAq_mean*LBXD05LAq+
          WLBXD07LAq_mean*LBXD07LAq+
WLBXF08LAq_mean*LBXF08LAq+ WLBXPCBLAq_mean*LBXPCBLAq+
WLBX044LAq_mean*LBX044LAq+
          WLBX049LAq_mean*LBX049LAq+
WLBX052LAq_mean*LBX052LAq+ WLBX087LAq_mean*LBX087LAq+
WLBX099LAq_mean*LBX099LAq+
          WLBX101LAq_mean*LBX101LAq+
WLBX110LAq_mean*LBX110LAq+ WLBX138LAq_mean*LBX138LAq+
WLBX146LAq_mean*LBX146LAq+
          WLBX149LAq_mean*LBX149LAq+
WLBX151LAq_mean*LBX151LAq+ WLBX153LAq_mean*LBX153LAq+
WLBX170LAq_mean*LBX170LAq+
          WLBX177LAq_mean*LBX177LAq+
WLBX178LAq_mean*LBX178LAq+ WLBX180LAq_mean*LBX180LAq+
WLBX183LAq_mean*LBX183LAq+
          WLBX187LAq_mean*LBX187LAq+
WLBX194LAq_mean*LBX194LAq+ WLBX196LAq_mean*LBX196LAq+
WLBX206LAq_mean*LBX206LAq+
          WLBX209LAq_mean*LBX209LAq+
WLBXBCDq_mean*LBXBCDq+ WLBXBPBq_mean*LBXBPBq+
WLBXTHGq_mean*LBXTHGq;
;

run;

```

*/

/*VALIDATE WEIGHTS FROM TEST DATASET*/

data together;

merge app.validate_small outstuff;

nutr= WDRACARq*DRTACARq+ WDRATOCq*DRTATOCq+
WDRBCARq*DRTBCARq+ WDRCAFFq*DRTCAFFq+ WDRTCALCq*DRTCALCq+
WDRTCARBq*DRTCARBq+
WDRTCOPPq*DRTCOPPq+ WDRTCRYPq*DRTC RYPq+ WDRTFaq*DRTFAq+
WDRTFDFEq*DRTFDFEq+
WDRTFFq*DRTFFq+ WDRTFIBEq*DRTFIBEq+
WDRTFOLAq*DRTFOLAq+ WDRTIROnq*DRTIROnq+ WDRTLyCOq*DRTLyCOq+
WDRTLZq*DRTLZq+ WDRTM161q*DRTM161q+
WDRTM181q*DRTM181q+ WDRTM201q*DRTM201q+ WDRTM221q*DRTM221q+
WDRTMAGNq*DRTMAGNq+
WDRTMFATq*DRTMFATq+ WDRTNIAcq*DRTNIAcq+ WDRTP182q*DRTTP182q+
WDRTP183q*DRTTP183q+
WDRTP204q*DRTTP204q+ WDRTP205q*DRTTP205q+
WDRTP225q*DRTTP225q+ WDRTP226q*DRTTP226q+ WDRTPFATq*DRTPFATq+
WDRTPHOSq*DRTPHOSq+
WDRTPOTAq*DRTPOTAq+ WDRTPROTq*DRTPROTq+ WDRTRETq*DRTRETq+
WDRTS040q*DRTS040q+
WDRTS060q*DRTS060q+ WDRTS080q*DRTS080q+
WDRTS100q*DRTS100q+ WDRTS120q*DRTS120q+ WDRTS140q*DRTS140q+
WDRTS160q*DRTS160q+ WDRTS180q*DRTS180q+
WDRtSELEq*DRTSELEq+ WDRtSFATq*DRTSFATq+ WDRtSODIq*DRTSODIq+
WDRtSUGRq*DRTSUGRq+
WDRtTFATq*DRTTFATq+ WDRtTHEOq*DRTTHEOq+ WDRtVARAq*DRTVARAq+
WDRtVB1q*DRTVB1q+
WDRtVB2q*DRTVB2q+ WDRtVB6q*DRTVB6q+
WDRtVB12q*DRTVB12q+ WDRtVCq*DRTVCq+ WDRtVKq*DRTVKq+
WDRtZINCq*DRTZINCq;

run;

/*

proc genmod data=together ;

class bin_race riagendr ;

model logalt =nutr bin_race age age*age riagendr indfmpir bmx bmi /type3;

ods output ParameterEstimates=parms;

run;

*/

```

/**/**/**/**/**/**/**/**/**/**/**/**/**/**/**/**BOOTSTRAP NUTRITION
WEIGHTS**/**/**/**/**/**/**/**/**/**/**/**/**/**/**/**/;

```

```

proc surveyselect data=app.test_small method=urs n=464
reps=1000 seed=113084 outhits out=work.boot_test;
run;

```

```

data start;
  _type_='PARMS';
  alpha=2; nutr=0.10001; CHEMS=0.1; b1=0.1; b2=0.1; b3=-0.2; b4=0.1; b5=0.01; B6=-0.1;
  sigma=1.0;
  array inwts

```

```

    WDRACARq  WDRATOCq  WDRBCARq  WDRCAFFq  WDRCALCq
WDRCARBq
    WDRCCOPPq  WDRCCRYPq  WDRCTFAq  WDRCTDFEq  WDRCTFFq  WDRCTFIBEq
WDRCTFOLAq  WDRCTIRONq
    WDRCTLYCOq  WDRCTLZq  WDRCTM161q  WDRCTM181q  WDRCTM201q  WDRCTM221q
WDRCTMAGNq  WDRCTMFATq  WDRCTNIACq
    WDRCTP182q  WDRCTP183q  WDRCTP204q  WDRCTP205q  WDRCTP225q  WDRCTP226q
WDRCTPFATq  WDRCTPHOSq
    WDRCTPOTAq  WDRCTPROTq  WDRCTRETq  WDRCTS040q  WDRCTS060q  WDRCTS080q
WDRCTS100q  WDRCTS120q  WDRCTS140q
    WDRCTS160q  WDRCTS180q  WDRCTSELEq  WDRCTSFATq  WDRCTSODIq
WDRCTSUGRq  WDRCTTFATq  WDRCTTHEOq  WDRCTVARAq
    WDRCTVB1q  WDRCTVB2q  WDRCTVB6q  WDRCTVB12q  WDRCTVCq  WDRCTVKq
WDRCTZINCq

```

```

;
```

```

  do over inwts;
    inwts=1/56;
  end;

```

```

proc nlp data=work.boot_test technique=trureg
  maxiter=10000 maxfunc=10000 inest=start outest=outstuff nomiss;

```

```

*set by variable for bootstrap samples;
by replicate;
*objective function;
  max logL;

```

```

*define parameters; *nutr;
  parms alpha nutr CHEMS b1 b2 b3 b4 b5 B6
    WDRACARq WDRATOCq WDRBCARq WDRCAFFq WDRCALCq
WDRCARBq
    WDRCOPPq WDRCRYPq WDRFAq WDRFDFFeq WDRFFq WDRFIBeq
WDRFOLAq WDRIRONq
    WDRLYCOq WDRTLZq WDRTM161q WDRTM181q WDRTM201q WDRTM221q
WDRTMAGNq WDRTMFATq WDRTNACq
    WDRTP182q WDRTP183q WDRTP204q WDRTP205q WDRTP225q WDRTP226q
WDRTPFATq WDRTPHOSq
    WDRTPOTAq WDRTPROTq WDRRETq WDRS040q WDRS060q WDRS080q
WDRS100q WDRS120q WDRS140q
    WDRS160q WDRS180q WDRTSELEq WDRTSFATq WDRTSODIq
WDRTSUGRq WDRTTFATq WDRTTHEOq WDRTVARAq
    WDRTVB1q WDRTVB2q WDRTVB6q WDRTVB12q WDRTVCq WDRTVKq
WDRTZINCq ;

```

```

*program statements;
  logL= -0.5*log(sigma)-0.5*(1/sigma)*(logALT-alpha-
- b1*bin_race - b2*RIDAGEYR -b6*RIDAGEYR*ridageyr- b3*riagendr- -b4*indfmpir-
b5*bmxbmi

```

```

-nutr* (WDRACARq*DRTACARq+ WDRATOCq*DRTATOCq+
WDRBCARq*DRTBCARq+ WDRCAFFq*DRTCAFFq+ WDRCALCq*DRTCALCq+
WDRCARBq*DRTCARBq+
WDRCOPPq*DRTCOPPq+ WDRCRYPq*DRTCRYPq+ WDRFAq*DRTFAq+
WDRFDFFeq*DRTFDFFeq+
WDRFFq*DRTFFq+ WDRFIBeq*DRTFIBeq+
WDRFOLAq*DRTFOLAq+ WDRIRONq*DRTIRONq+ WDRLYCOq*DRTLYCOq+
WDRTLZq*DRTLZq+ WDRTM161q*DRTM161q+
WDRTM181q*DRTM181q+ WDRTM201q*DRTM201q+ WDRTM221q*DRTM221q+
WDRTMAGNq*DRTMAGNq+
WDRTMFATq*DRTMFATq+ WDRTNACq*DRTNACq+ WDRTP182q*DRTTP182q+
WDRTP183q*DRTTP183q+
WDRTP204q*DRTTP204q+ WDRTP205q*DRTTP205q+
WDRTP225q*DRTTP225q+ WDRTP226q*DRTTP226q+ WDRTPFATq*DRTTPFATq+
WDRTPHOSq*DRTTPHOSq+
WDRTPOTAq*DRTTPOTAq+ WDRTPROTq*DRTTPROTq+ WDRRETq*DRTRETq+
WDRS040q*DRTS040q+
WDRS060q*DRTS060q+ WDRS080q*DRTS080q+
WDRS100q*DRTS100q+ WDRS120q*DRTS120q+ WDRS140q*DRTS140q+
WDRS160q*DRTS160q+ WDRS180q*DRTS180q+
WDRTSELEq*DRTSELEq+ WDRTSFATq*DRTSFATq+ WDRTSODIq*DRTSODIq+
WDRTSUGRq*DRTSUGRq+
WDRTTFATq*DRTTFATq+ WDRTTHEOq*DRTTHEOq+ WDRTVARAq*DRTTVARAq+
WDRTVB1q*DRTVB1q+

```

WDRTVB2q*WDRTVB2q+ WDRTVB6q*WDRTVB6q+
WDRTVB12q*WDRTVB12q+ WDRTVCq*WDRTVCq+ WDRTVKq*WDRTVKq+
WDRTZINCq*WDRTZINCq))**2;

*linear constraints;

lincon

/*weights for Nutrients sum to 1*/

WDRTACARq+ WDRTATOCq+ WDRTBCARq+
WDRTCAFFq+ WDRTCALCq+
WDRTCARBq+ WDRTCOPPq+ WDRTCRYPq+
WDRTFAq+ WDRTFDfEq+
WDRTFFq+ WDRTFIBEq+ WDRTFOLAq+
WDRTIRONq+ WDRTLYCOq+
WDRTLZq+ WDRTM161q+ WDRTM181q+
WDRTM201q+ WDRTM221q+
WDRTMAGNq+ WDRTMFATq+ WDRTNIACq+
WDRTP182q+ WDRTP183q+
WDRTP204q+ WDRTP205q+ WDRTP225q+
WDRTP226q+ WDRTPFATq+
WDRTPHOSq+ WDRTPOTAq+ WDRTPROTq+
WDRTRETq+ WDRTS040q+
WDRTS060q+ WDRTS080q+ WDRTS100q+
WDRTS120q+ WDRTS140q+
WDRTS160q+ WDRTS180q+ WDRTSELEq+
WDRTSFATq+ WDRTSODIq+
WDRTSUGRq+ WDRTTFATq+ WDRTTHEOq+
WDRTVARAq+ WDRTVB1q+
WDRTVB2q+ WDRTVB6q+ WDRTVB12q+
WDRTVCq+ WDRTVKq+ WDRTZINCq=1;

*bounds on weights: all in (0,1);

bounds

0<WDRTACARq<1, 0<WDRTATOCq<1,
0<WDRTBCARq<1, 0<WDRTCAFFq<1, 0<WDRTCALCq<1,
0<WDRTCARBq<1, 0<WDRTCOPPq<1,
0<WDRTCRYPq<1, 0<WDRTFAq<1, 0<WDRTFDfEq<1,
0<WDRTFFq<1, 0<WDRTFIBEq<1, 0<WDRTFOLAq<1,
0<WDRTIRONq<1, 0<WDRTLYCOq<1,
0<WDRTLZq<1, 0<WDRTM161q<1, 0<WDRTM181q<1,
0<WDRTM201q<1, 0<WDRTM221q<1,
0<WDRTMAGNq<1, 0<WDRTMFATq<1,
0<WDRTNIACq<1, 0<WDRTP182q<1, 0<WDRTP183q<1,
0<WDRTP204q<1, 0<WDRTP205q<1,
0<WDRTP225q<1, 0<WDRTP226q<1, 0<WDRTPFATq<1,

```

0<WDRTPHOSq<1, 0<WD RTPOTAq<1,
0<WD RTPROTq<1, 0<WD RTRETq<1, 0<WD RTS040q<1,
0<WD RTS060q<1, 0<WD RTS080q<1,
0<WD RTS100q<1, 0<WD RTS120q<1, 0<WD RTS140q<1,
0<WD RTS160q<1, 0<WD RTS180q<1,
0<WD RTSELEq<1, 0<WD RTSFATq<1, 0<WD RTSODIq<1,
0<WD RTSUGRq<1, 0<WD RTTFATq<1,
0<WD RTTHEOq<1, 0<WD RTVARAq<1, 0<WD RTVB1q<1,
0<WD RTVB2q<1, 0<WD RTVB6q<1,
0<WD RTVB12q<1, 0<WD RTVCq<1, 0<WD RTVKq<1, 0<WD RTZINCq<1,

```

```

/*determine set of "protective" nutrients by constraining vits to be negative and "negative" chems
by constraining pcbs positive*/

```

```

    nutr>0.05;

```

```

run;

```

```

data work.outstuff;
set work.outstuff;
where _type_='PARMS';
run;

```

```

/*Validate bootstrap samples in Validate dataset*/
/*

```

```

data replicates;
do replicate=1 to 1000;
    do obs=1 to 464; output;
    end;
end;
run;

```

```

data valid_rep;
set app.validate_small;
obs=_n_;
run;
proc sort data=work.valid_rep;
by obs;
run;

```

```

proc sort data=work.replicates;
by obs;
run;

```

```

data valid_rep_tog;
merge valid_rep replicates;

```

```

by obs;
run;

proc sort data=valid_rep_tog;;
by replicate;
run;

proc sort data=outstuff;
by replicate;
run;
*/

/*VALIDATE WEIGHTS FROM TEST DATASET*/
data together;
merge vboot_test outstuff; by replicate;

    nutr=
        WDR TACARq*DRTACARq+ WDR TATOCq*DRTATOCq+
WDR TBCARq*DRTBCARq+ WDR TCAFFq*DRTCAFFq+ WDR TCALCq*DRTCALCq+
        WDR TCARBq*DRTCARBq+
        WDR TCOPPq*DRTCOPPq+ WDR TCRYPq*DTRTCRYPq+ WDR TFAq*DRTFAq+
WDR TFD FEq*DRTFD FEq+
        WDR TFFq*DRTFFq+ WDR TFIBEq*DRTFIBEq+
WDR TFOL Aq*DRTFOL Aq+ WDR TIRONq*DRTIRONq+ WDR TLYCOq*DRTLYCOq+
        WDR TLZq*DRTLZq+ WDR TM161q*DRTM161q+
WDR TM181q*DRTM181q+ WDR TM201q*DRTM201q+ WDR TM221q*DRTM221q+
        WDR TMAGNq*DRTMAGNq+
WDR TMFATq*DRTMFATq+ WDR TNIACq*DRTNIACq+        WDR TP182q*DRTTP182q+
WDR TP183q*DRTTP183q+
        WDR TP204q*DRTTP204q+ WDR TP205q*DRTTP205q+
WDR TP225q*DRTTP225q+ WDR TP226q*DRTTP226q+ WDR TPFATq*DRTPFATq+
        WDR TPHOSq*DRTPHOSq+
        WDR TPOTAq*DRTTPOTAq+ WDR TPROTq*DRTPROTq+ WDR TRETq*DRTRETq+
WDR TS040q*DRTS040q+
        WDR TS060q*DRTS060q+ WDR TS080q*DRTS080q+
WDR TS100q*DRTS100q+ WDR TS120q*DRTS120q+ WDR TS140q*DRTS140q+
        WDR TS160q*DRTS160q+ WDR TS180q*DRTS180q+
WDR TSELEq*DRTSELEq+ WDR TSFATq*DRTSFATq+ WDR TSODIq*DRTSODIq+
        WDR TSUGRq*DRTSUGRq+
WDR TTFFATq*DRTTFFATq+ WDR TTHEOq*DRTTHEOq+ WDR TVARAq*DRTVARAq+
        WDR TVB1q*DRTVB1q+
        WDR TVB2q*DRTVB2q+ WDR TVB6q*DRTVB6q+
WDR TVB12q*DRTVB12q+ WDR TVCq*DRTVCq+ WDR TVKq*DRTVKq+
WDR TZINCq*DRTZINCq;

run;

data together;
set work.together;

```

```

if ridreth1=3 then bin_race=1;
else bin_race=0;
run;

```

```

proc genmod data=together ;
by replicate;
class bin_race riagendr ;
  model logalt = nutr bin_race RIDAGEYR RIDAGEYR*RIDAGEYR riagendr indfmpir
  bmx bmi/type3;/*nutr*/
  ods output ParameterEstimates=parms;
run;

```

```

data nutr;
set work.parms;
where parameter='nutr';
if probchisq<=0.05 then power=1;
else power=0;
run;

```

```

proc freq data=nutr;
tables power;
run;

```

```

data nutr_weights;
merge nutr outstuff;
by replicate;
run;

```

```

ods rtf file='C:\Users\carrck\Documents\Dissertation\application chapter\Application
Chapter\means_nutr.rtf';
proc means data=nutr_weights;
where power=1;
var WDR TACARq WDR TATOCq WDR TBCARq WDR TCAFFq WDR TCAFCq WDR TCAFCq
WDR TCARBq
WDR TCOPPCq WDR TCRCYPq WDR TFACq WDR TFDFFeq WDR TFFeq WDR TFIBEq
WDR TFOFACq WDR TIRONq
WDR TLYCOq WDR TLZq WDR TM161q WDR TM181q WDR TM201q WDR TM221q
WDR TMAGNq WDR TMFATq WDR TNIAcq
WDR TP182q WDR TP183q WDR TP204q WDR TP205q WDR TP225q WDR TP226q
WDR TPFAcq WDR TPPOSq
WDR TPOTACq WDR TPROTq WDR TRETCq WDR TS040q WDR TS060q WDR TS080q
WDR TS100q WDR TS120q WDR TS140q
WDR TS160q WDR TS180q WDR TSELEq WDR TSFATq WDR TSODIq
WDR TSUGRq WDR TTFATq WDR TTTHOq WDR TVARACq
WDR TVB1q WDR TVB2q WDR TVB6q WDR TVB12q WDR TVCq WDR TVKq
WDR TZINCq;

```



```
ods output summary=means;
```

```
run;  
ods rtf close;
```

```
data validate;  
set app.validate_small;  
dummy=1;  
run;
```

```
data means;  
set work.means;  
dummy=1;  
run;
```

```
data val_means;  
merge validate means;  
by dummy;  
run;
```

```
data val_means;  
set val_means;  
nutr_mean=      WDR TACARq_mean*DRTACARq+  
WDR TATOCq_mean*DRTATOCq+ WDR TBCARq_mean*DRTBCARq+  
WDR TC AFFq_mean*DRTC AFFq+ WDR TCA LCq_mean*DRTCA LCq+  
                WDR TCARBq_mean*DRTCARBq+  
WDR TCOPPq_mean*DRTCOPPq+ WDR TCRYPq_mean*DRTC RYPq+  
WDR TF Aq_mean*DRTFAq+ WDR TFDFEq_mean*DRTFDFEq+  
                WDR TFFq_mean*DRTFFq+  
WDR TFIBEq_mean*DRTFIBEq+ WDR TFOLAq_mean*DRTFOLAq+  
WDR TI RONq_mean*DRTI RONq+ WDR TLYCOq_mean*DRTLYCOq+  
                WDR TLZq_mean*DRTLZq+  
WDR TM161q_mean*DRTM161q+ WDR TM181q_mean*DRTM181q+  
WDR TM201q_mean*DRTM201q+ WDR TM221q_mean*DRTM221q+  
                WDR TMAGNq_mean*DRTMAGNq+  
WDR TMFATq_mean*DRTMFATq+ WDR TNIA Cq_mean*DRTNIA Cq+  
                WDR TP182q_mean*DRTP182q+ WDR TP183q_mean*DRTP183q+  
                WDR TP204q_mean*DRTP204q+  
WDR TP205q_mean*DRTP205q+ WDR TP225q_mean*DRTP225q+  
WDR TP226q_mean*DRTP226q+ WDR TPFA Tq_mean*DRTPFA Tq+  
                WDR TPHOSq_mean*DRTPHOSq+  
                WDR TPOTAq_mean*DRTPOTAq+ WDR TPROTq_mean*DRTPRO Tq+  
WDR TRE Tq_mean*DRTRE Tq+ WDR TS040q_mean*DRTS040q+
```

```

WDRTS060q_mean*DRTS060q+
WDRTS080q_mean*DRTS080q+ WDRTS100q_mean*DRTS100q+
WDRTS120q_mean*DRTS120q+ WDRTS140q_mean*DRTS140q+
WDRTS160q_mean*DRTS160q+
WDRTS180q_mean*DRTS180q+ WDRTSELEq_mean*DRTSELEq+
WDRTSFATq_mean*DRTSFATq+ WDRTSODIq_mean*DRTSODIq+
WDRTSUGRq_mean*DRTSUGRq+
WDRTTFATq_mean*DRTTFATq+ WDRTTHEOq_mean*DRTTHEOq+
WDRTVARAq_mean*DRTVARAq+ WDRTVB1q_mean*DRTVB1q+
WDRTVB2q_mean*DRTVB2q+
WDRTVB6q_mean*DRTVB6q+ WDRTVB12q_mean*DRTVB12q+
WDRTVCq_mean*DRTVCq+ WDRTVKq_mean*DRTVKq+
WDRTZINCq_mean*DRTZINCq;
;

```

```
run;
```

```

proc genmod data=val_means;
class bin_race riagendr ; *nutr_mean;
model logalt = nutr_mean bin_race RIDAGEYR RIDAGEYR*RIDAGEYR riagendr
indfmpir bmx bmi/type3;
ods output ParameterEstimates=parms;
run;

```

```

data app.parms_nutr0304;
set work.parms; run;

```

```

data app.val_means;
merge val_means app.val_means;
by seqn;
run;

```

```
proc contents data=app.val_means; run;
```

```

proc genmod data=app.val_means;
class bin_race riagendr ;
model logalt = nutr_mean chems_mean bin_race RIDAGEYR RIDAGEYR*RIDAGEYR
riagendr indfmpir bmx bmi/type3;
ods output ParameterEstimates=parms;
run;

```

```
data app.FINAL;
```

```

set app.val_means;
AGE=RIDAGEYR/10;
run;

```

```

proc genmod data=app.val_means;
class bin_race riagendr ;
  model logalt = nutr_mean chems_mean bin_race RIDAGEYR RIDAGEYR*RIDAGEYR
riagendr indfmpir bmx bmi/type3;
  ods output ParameterEstimates=parms;
run;

```

```

libname app 'C:\Users\carrrck\Documents\Dissertation\application chapter\Application Chapter';

```

```

data app.final;
set app.final;
age=ridageyr/10;
run;

```

```

symbol v=dot i=sm80s;
proc gplot data=app.final;
plot logalt*ridageyr;
label ridageyr= 'Age in Years';
run;

```

```

ods html newfile=proc;
proc genmod data=app.final_mar05;
class bin_race riagendr ;
  model logalt = chems_mean bin_race age age*age riagendr indfmpir bmx bmi/type3;
  ods output ParameterEstimates=parms;
run;

```

```

proc genmod data=app.final_mar05;
class bin_race riagendr ;
  model logalt = nutr_mean age age*age bin_race riagendr indfmpir bmx bmi/type3;
  ods output ParameterEstimates=parms;
run;

```

```

proc genmod data=app.final_mar05;
class bin_race riagendr ;
  model logalt = chems_mean nutr_mean age age*age bin_race riagendr indfmpir
bmx bmi/type3;
  ods output ParameterEstimates=parms;
run;

```

```

proc genmod data=app.final_mar05;
class bin_race riagendr ;
  model logalt = chems_mean nutr_mean chems_mean*nutr_mean age age*age bin_race
riagendr indfmpir bmx bmi/type3;
  ods output ParameterEstimates=parms;
               run;

```

```

/*Plot Surface*/
data app.pcb_nutr_data_nomiss;
set app.pcb_nutr_data_nomiss;
if ridreth1=3 then bin_race=1;
else bin_race=0;
age=ridageyr/10; run;

```

```

proc means data=app.pcb_nutr_data_nomiss;
var bin_race age riagendr indfmpir bmx bmi;
run;

```

```

data forplot;
do ECS=0 to 3 by 0.1;
do NSS=0 to 3 by 0.1; output;
end; end; run;

```

```

data app.forplot;
set work.forplot;
alt_avg= exp(2.516+ECS*0.091 +NSS*0.132);
alt_men=exp(2.863+ECS*0.091 +NSS*0.132);
alt_women=exp(2.637+ECS*0.091 +NSS*0.132);
; run;

```

```

proc means data=app.forplot;
var alt_men;

```

```

run;
goptions htext=1.7; run;
proc g3d data=app.forplot;
plot ECS*NSS=alt /grid;
label alt= 'Mean ALT';
run;

```

```

ods rtf file='C:\Users\carrck\Documents\Dissertation\application chapter\Application
Chapter\contours.rtf';
goptions reset=all;
goptions htext=1.7 font=swiss;
axis1 LABEL=(angle=90 COLOR=black "NSS");
symbol1 Value="18"
        color=black

```

```

        height=1.2
            font="swiss";
symbol2 Value="20"
        color=black
        height=1.2
            font="swiss";
symbol3 Value="22"
        color=black
        height=1.2
            font="swiss";
symbol4 Value="24"
        color=black
        height=1.2
            font="swiss";
symbol5 Value="26"
        color=black
        height=1.2
            font="swiss";
symbol6 Value="28"
        color=black
        height=1.2
            font="swiss";
symbol7 Value="RISK(30)"
        color=red
        height=1.5
            font="swiss";
symbol8 Value="RISK(32)"
        color=red
        height=1.5
            font="swiss";
proc gcontour data=app.forplot;
plot NSS*ECS=alt_men /autolabel=(check=none) levels= 18 to 32 by 2 vaxis=axis1 vref=2
wvref=3 nolegend;
label alt_men="Predicted Mean ALT for Men";
run;

goptions reset=all;
goptions htext=1.7 font=swiss;
axis1 LABEL=(angle=90 COLOR=black "NSS");
symbol1 Value="15"
        color=black
        height=1.5
            font="swiss";
symbol2 Value="17"
        color=black
        height=1.5
            font="swiss";
symbol3 Value="RISK(19)"

```

```

        color=red
    height=1.5
        font="swiss";
symbol4 Value="RISK(21)"
        color=red
    height=1.5
        font="swiss";
symbol5 Value="RISK(23)"
        color=red
    height=1.5
        font="swiss";
symbol6 Value="RISK(25)"
        color=red
    height=1.5
        font="swiss";
symbol7 Value="RISK(27)"
        color=red
    height=1.5
        font="swiss";
symbol8 Value="RISK(29)"
        color=red
        height=1.5
        font="swiss";
        goptions htext=1.7 font=swiss;

proc gcontour data=app.forplot ;
plot NSS*ECS=alt_women /autolabel=(check=none) levels= 15 to 27 by 2 vaxis=axis1
vref=0.25 wvref=3 nolegend;
label alt_women= "Predicted Mean ALT for Women";
run;

ods rtf close;
proc print data=app.chems_means0304;
run;

```