Interuniversity Institute for Biostatistics and Statistics and Statistical Bioinformatics

Universiteit Hasselt

Human Biomonitoring: Multicollinearity

Mekdes Abera Taye

Internal Supervisor: Liesbeth Bruckers (Msc)

Thesis submitted in partial fulfillment of the requirements for the Degree of Master of Science in Epidemiology and Public health Methodology

September 11, 2013

Contents

Acknowledgments

Thanks to the Almighty God who sustained me to this status and for his unlimited blessings. Next, I am very grateful to, my internal supervisor Mrs. Liesbeth Bruckers, for her guidance, suggestions, comments and willingness to help me.

I would like to take this opportunity to give my heartfelt appreciation to all professors, instructors and staff member specially Mrs. Martin at Uhasselt for sharing their knowledge and skills.

I am also great full to Mr, Lioneal Duncan for his prays, and kind cooperation on providing material and moral supports. His advices were my strengths on those days I lost hope. God bless you more and more Mr. Lionel.

Samiya, I don't know what to say for your continued support, prayers, encouragement and love from the day we met until today and for the coming future.

Finally I want to say thanks to my father (Mr. Abera Taye) and my mother (Mrs. Selamnesh Alemu), for continual prayer, moral support, constructive advices, encouragement and for their unconditional love.

Abstract

Biomonitoring is a scientific technique that assesses a person's exposure to natural and synthetic chemicals. The data from biomonitoring program reveal the amounts of natural and manufactured chemicals that entered human body (bio-marker of exposure) and their effect on human health. In this paper we investigate the influence of a number of covariates (levels of pollutants, gender, life style...) on Thyroid Stimulated Hormone. The covariates, however, had strong correlation ship, which hamper the efficiency of simple OLS as technique of estimation. Thus, we used ridge, lasso and elastic net regression methods as technique of estimation and tested their performance in inclusion of highly correlated covariates in the model. We obtained that all regularization methods improve the prediction accuracy as compared to no shrinking (OLS). Among the regularization methods, Elastic Net performs better than Lasso, and the lasso performs better than ridge regression in terms of prediction accuracy.

Key words: Ridge, Lasso, Elastic Net, Cross-Validation.

Abbreviations

DDE: Dichlorodiphenyldichloro Ethylene
HCB: Hexachlorobenzene
PCBs: Polychlorinated biphenyls
HBA: Paraben Metabolite
BDEs: Brominated Diphenylethers
HBCD: hexabromocyclododecane
TRA: Toxic relevant Arrsene
DCP: Dichloropenol
MBZP: Mono benzyl Phthalate
DEP: Diethylphosphate
HBA: Para-Hydroxybenzoic Acid
MBC: 4-Methylbenzylidene Camphor
BEP: Persistent Brominated Pollutants
HCB: Hexachlorobenzene
SOMPCB: sum of PCB118 and PCB170
SOMME: sum of MEHP(mono-2-ethylhexyl ftalaat), MEHHP(mono-2-ethylhexyl-5-hydroxyhexyl ftalaat) and MEOHP(mono-2-ethyl-5-oxohexyl ftalaat)
UTCS: Urinary triclosan
COT: Cotinine
UTL: Urinary Thallium Concentration
UCU: Urinary Copper Concentration
UCD: Urinary Cadmium Concentration
UBPA: Urinary Bisphenol A Concentration
UAS: Urinary Arsenic Concentration
TRA: Toxic Relevant Arsenic
HPYR: 1-hydroxypyreen
MBP: Monobutyl Phthalate
MBzP: Mono Benzyl Phthalate
TTMA: t,t'-muconzuur
PCDD_F: Polychlorodibenzodioxins Fluoro
DL_PCB: Dioxin-Like Polychlorinated Bipheny

1. Introduction

Biomonitoring is a scientific technique that assesses a person's exposure to natural and synthetic chemicals. In biomonitoring studies, scientists examine and measure the concentration of chemicals in a person's tissues and fluids, most commonly blood, urine, breast milk and sometimes expelled air. Since chemicals that entered human body leave markers, scientists can examine the level of exposure through detecting the chemical itself or a breakdown product, or a change in the body resulting from the interaction of the chemical or its breakdown product(s) with the individual, such as alterations in the levels of certain enzymes or other proteins which may lead to modifications of normal body processes.

The process of Human Biomonitoring involves three key initial steps in: selecting who will be monitored, as well as when and where, collecting tissue and/or fluid samples, and deciding which environmental pollutants to study and analyzing those environmental pollutants in the samples that are collected. This is a complex processes that require sophisticated analytical instruments and methods.

The data from biomonitoring reveal the amounts of natural and manufactured chemicals that have entered and remain in the body by age, sex, geographical location, and state of health of the individual. This information can give an indication into sources of chemical exposure, possible health effects, and ways to reduce or prevent future exposure. It is, therefore, helpful to develop new political strategies, to adapt existing political strategies and/or to evaluate and assesses existing political strategies.

Since 2002 a human biomonitoring network has been established in Flanders (Belgium) as part of a program on environmental health surveillance. Biomonitoring research has been carried out by different partners of the Centre for Environment and Health: the Flemish Institute for Technological Research (VITO), the Provincial Institute of Hygiene (PIH) and all Flemish universities (Brussels, Antwerp, Ghent, Hasselt and Leuven) with the aim of providing better insights in the impact of the environment on human health, and providing a profound base for a more effective environment and health policy.

In the period 2002-2006 the Flemish government initiated a five-year first generation human biomonitoring programme carried out by Flemish Centre of Expertise on Environment and Health. During this time a biomonitoring study was conducted in newborns, adolescents and adults in various areas (Albert Canal, port areas, incinerators, fruit area etc) with a different environmental load. This measurement campaign confirmed that living in different regions in Flanders can be associated with pollutants in the body and its impact on health.

After a first programme period, the centre continued the programme for the next five years (2007–2011); expanding substantially the number of environmental chemicals for which human biomonitoring data were obtained. The studies were conducted among young people, new-born and adults in the industrial areas of the Genk-Zuid and the region Menen. The hypothesis is tested whether in the hot spot areas specific biomonitoring data (exposure and effects) are different from reference values that have been obtained over Flanders.

For the period 2012-2015, a new program launched to build further on the strengths and knowledge gained in the 2002-2006 and 2007-2011 programs. The measurement campaign launched in areas known for its industrial activities, busy road and ship traffic and the associated emissions issue among young people in the Ghent canal on which this study is a part. The main focus of the Flemish biomonitoring program remains to use and translate the scientific results into policy actions.

1.1 Statement of the Problem

As reported in the literature, many chemicals released into the environment can disturb the development of the endocrine system and of the organs that respond to endocrine signals and can interfere with the correct functioning of endocrine organs. Such chemicals are generally named “Endocrine Disruptors” (EDs). Persistent organic pollutants (POPs) are known Endocrine Disrupting Chemicals (EDCs). Persistent organic pollutants (POPs) are chemical substances that persist in the environment, bioaccumulate through the food, and pose a risk of causing adverse effects to human health and the environment. They include polychlorinated biphenyls (PCBs) and organochlorine pesticides, such as dioxin-like PCBs (PCB118), non-dioxin like PCBs (PCB138, 153, 170, and 180), dichlorodiphenyldichloroethylene (DDE) and hexachlorobenzene (HCB). Dioxin-like PCBs exhibit similar toxicity as dioxins. Dioxin-like PCBs are formed by

incomplete burning processes, in steel industry, exhaust fumes, and cigarette smoke. Humans are exposed via fat fish, whole milk, and milk products, fat meat and products containing animal fats (SCQRF, 2000). The non-dioxin like PCBs were often used in industrial and commercial applications, such as hydraulic systems, cooling fluids, coloring dyes, and in plastic and rubber products. Production of PCBs has been banned in most developed countries since late 1980s (EFSA, 2005). However, PCBs are still released in the environment by incorrect removal and leakage in electrical devices and hydraulic systems. Humans are exposed to PCBs in food especially via fish, but also via meat and animal products. Fruit, vegetables, grain and grain products are less important as sources for PCBs (EFSA, 2005). Dichlorodiphenyltrichloroethane (DDT), the metabolite of DDE, is an insecticide used in agriculture. It is present in tuberous plants, green vegetables, and in fat meat, fish and chicken (EFSA, 2006a). HCB is an anti-fungal agent that was used previously to protect seeds and grains. It is also an industrial product, used in the past in the production of fireworks, munitions and synthetic rubber. Humans are exposed by food via fat fish, whole milk, whole milk products and fat meat (EFSA, 2006b).

Polybrominated diphenyl ethers (PBDEs) are bioaccumulative brominated flame retardants currently ubiquitous in the environment (Hites, 2004; Hoh and Hites, 2005). They are additives in some plastics, foams, electronics, and fabrics, originating from three commer-cial mixtures: penta-, octa-, and deca-BDE.

These chemical groups all have different mechanisms of endocrine action, and some evidence exists of their effects on reproductive, developmental and other homeostatic systems by interfering with synthesis, secretion, transport, metabolism and action of endogenous hormones. Many such chemicals also affect the thyroid hormone (TH) system at multiple levels. Thyroid hormone is essential for the control of metabolism, normal brain development, and for many aspects of normal adult physiology. Changes in the function of the thyroid gland or interference with the ability of thyroid hormone to exert its action may produce undesired effects on development, metabolism, or adult physiology. Brucker-Davis (1998) and Howdeshell (2002) have extensively reviewed a large number of industrial chemicals have been found to alter circulating levels of thyroid hormone. Diamanti-Kandarakis et al. (2009) also provided a brief overview of the literature regarding the mechanisms by which environmental chemicals may interfere with thyroid hormone action, with a brief background of thyroid endocrinology.

The present study aims at assessing the association of a number of pollutants (bio-markers of exposure) and other covariates (such as life style and education level) with thyroid stimulating hormone (TSH). On the study, however, we contend with the problem of estimating the independent effects of many correlated exposures. General approaches in this regard include assessing each exposure separately, adjusting for some subset of other exposures, or assessing all exposures simultaneously in a single model. The optimal strategy remains uncertain, and it is unclear to what extent different reasonable approaches influence findings. On the study our scope, however, is not provide an empirical comparison of results from several modeling strategies, but is to investigate if shrinkage methods such as Ridge, Lasso and Elastic net are valuable techniques that allow inclusion of highly correlated covariates simultaneously in the model.

The report is organized as follows. Section 2 and 3 describes the data and the statistical methods implemented in the analysis respectively. The results of the analysis are presented in section 4 while section 5 presents a brief discussion and conclusion of the analysis. The references and appendix are presented in sections 6 and 7 respectively.

1.2 Objectives of the Study

The main objective of the study is to investigate the influence of a number of covariates (levels of pollutants, gender, and education) on effect biomarkers (Thyroid stimulated hormone). Ordinary least square (OLS) regression model is one of the techniques used to identify the risk factors. This method does poorly when strongly correlated predictors tend to be in the model together. It tends to affect the inferences about the mean response or prediction of new observation. Penalization techniques have been proposed to improve OLS which include techniques like elastic net regression, ridge regression and lasso. We employ these techniques and test if they are valuable to solve multicollinearity problem in our case and make comparison between these methods.

2. Dataset

Since 2002, the Flemish human biomonitoring program have been collecting data on a variety of biomarkers of exposure (including e.g. DDE, heavy metals, dioxins, exposure to PAHs and benzene) and effect (including e.g. DNA damage, asthma and allergy, TSH, etc.) in different

areas of Flanders for three age groups (newborn babies, adolescents, elderly). A stratified clustered multi- stage design was used to select the respondents as a random sample of the population residing in Flanders. Different questionnaires and laboratory analysis were used to obtain information about the study population background and covariates. The present study uses recent data collected on measurement campaign launched in areas known for its industrial activities, busy road and ship traffic and the associated emissions issue among adolescents (aged 14 - 15) in the Flanders. We have a total of 210 observations with 35 covariates. After deletion of the 14 observations with missing values for categorical covariates, we used the remaining 196 observations in our analysis.

3. Statistical Methods

This section presents a description of the statistical methods used in order to meet the objective of the study. Besides explanatory data analysis techniques, we widely used ridge regression, lasso, and elastic net which are briefly discussed below. SAS 9.2 and R 2.15 are the main softwares used in this analysis.

3.1 Exploratory Data Analysis

Exploratory data analysis (EDA) - summary statistics (mean, median and standard deviation), pair-wise correlation and graphs were applied to gain an insight into the dataset.

3.2. Ridge Regression

OLS regression may result in highly variable estimates of the regression coefficients in the presence of multicollinearity or when the number of predictors (p) is large relative to the number of observations(n). Ridge regression (Hoerl and Kennard, 1988) reduces this variability by shrinking the coefficients, resulting in more prediction accuracy at the cost of usually only a small increase of bias. It minimizes the residual sum of squares subject to a bound on the ℓ_2 norm of the coefficients. In Ridge regression, the coefficients are shrunken towards zero, but will never become exactly zero. The ridge regression estimator solves the regression problem using ℓ_2 penalized least squares:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{ridge}) = \underset{\beta}{argmin}\left|\left|\boldsymbol{y} - \boldsymbol{X\beta}\right|\right|_2^2 + \lambda_2\left|\left|\boldsymbol{\beta}\right|\right|_2^2$$

Where $\left|\left|y - \mathbf{X\beta}\right|\right|_2^2 = \sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2$ is the ℓ_2-norm (quadratic) loss function (i.e. residual sum of squares), x_i^T is the i^{th} patient (row of *X*), $\left|\left|\boldsymbol{\beta}\right|\right|_2^2 = \sum_{j=1}^{p}\beta_j^2$ is the ℓ_2 –norm penalty on β, and $\lambda_2 \geq 0$ is the tuning (penalty or complexity) parameter which regulates the strength of the penalty (linear shrinkage). The larger the value of λ_2, the greater the amount of shrinkage. So, when the number of predictors is large, ridge regression will not provide a sparse model that is easy to interpret.

The general trend is:

- The bias increases as λ_2 (amount of shrinkage) increases.
- The variance decreases as λ_2 (amount of shrinkage) increases.

As a continuous shrinkage method, ridge regression achieves its better prediction performance through a bias–variance trade-off (decrease in variance for small increase of bias). However, ridge regression cannot produce a parsimonious model, for it always keeps all the predictors in the model. The solution of ridge regression problem with parameter λ_2 is given by:

$$\hat{\beta}_{ridge} = (\boldsymbol{X'X} + \lambda_2\boldsymbol{I})^{-1}\boldsymbol{X'}Y$$

Where, **I** is pxp identity matrix. The method of ridge is implemented in R using the package *MASS*.

3.3. Least Absolute Shrinkage and Selection Operator (LASSO)

A promising technique called the lasso was proposed by Tibshirani (1996). It is a penalized least squares method imposing an ℓ_1-penalty on the regression coefficients. The lasso is, however, not robust to high correlations among predictors and will arbitrarily choose one and ignore the others and break down when all predictors are identical (Friedman, Hastie and Tibshirani, 2010).

Lasso does both continuous shrinkage and automatic variable selection simultaneously. It reduces the variability of the estimates by shrinking the coefficients and at the same time produces interpretable models by shrinking some coefficients to exactly zero (see fig 1 below). The lasso estimator uses the ℓ_1 penalized least squares criterion to obtain a sparse solution to the following optimization problem (Tibshirani, 1996):

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{Lasso}) = \underset{\beta}{argmin}\left|\left|\boldsymbol{y} - \boldsymbol{X\beta}\right|\right|_2^2 + \lambda_1\left|\left|\boldsymbol{\beta}\right|\right|_1$$

Where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 –norm penalty on **β**, which induces sparsity in the solution, and $\lambda_1 \geq 0$ is a tuning parameter. The regression coefficients estimated as:

$$\boldsymbol{\beta}^{lasso} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\boldsymbol{X}'y - \frac{\lambda_1}{2}w\right)$$

Where, the elements w_j of **w** are either +1 or −1, depending on the sign of the corresponding regression coefficient β_j.

Even so, the lasso has key shortcomings, it is unstable with high-dimensional data, cannot select more predictors (p) than the sample size (n) if $p > n$ and it selects one member of a set of highly correlated predictors. The method of lasso is implemented in R using the package *lasso2*.

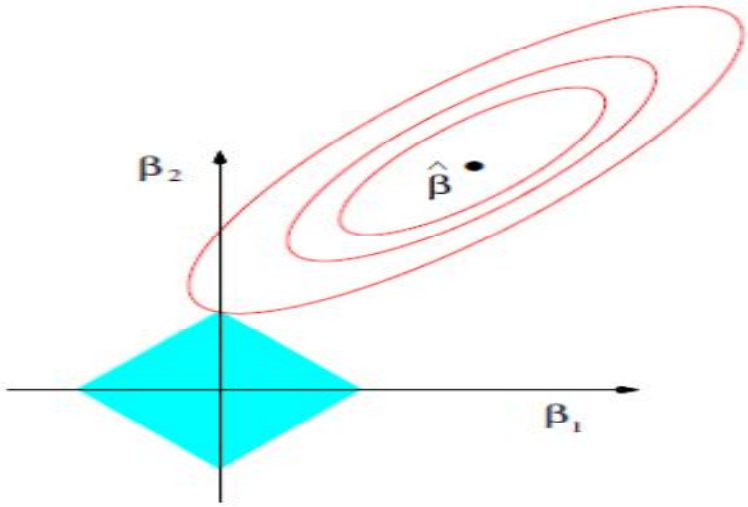

Figure 1: plot for the lasso how works from Tibshirani, 1996. Show is contour of the error and the constraint function. The solid aqua shaded areas are the constraint regions $|\beta_1|+|\beta_1|\leq t^2$*, while the red ellipses are the contours of the least square error function.*

Figure 1 illustrates how lasso works, in the special case of 2 predictors. The ellipses are contours of constant residual sum of squares (RSS), which is minimized at the point marked $\hat{\beta}$. As you move away from $\hat{\beta}$ the residual sum of squares increases, but all points on the same elliptical contour have the same value of RSS. The aqua shaded square is the set of vectors β that satisfy the constraint $\sum|\beta| \leq \lambda$ and for given λ the estimator will be the point with smallest residual sum of squares, and so the lasso estimator will be the point in the aqua square, or clearly on the boundary of the aqua shaded square unless $\hat{\beta} = 0$, that lies on the contour closest to $\hat{\beta}$.

3.4. The Elastic Net

Zou and Hastie (2005) proposed the Elastic Net to overcome the limitations of the Lasso and Ridge regression in some situations. The elastic net (ENET) is an extension of the lasso that is robust to extreme correlations among the predictors (Friedman, Hastie and Tibshirani, 2010). It also combines shrinkage and variable selection, and in addition encourages grouping of variables: groups of highly correlated variables tend to be selected together, where as Lasso would only select one variable of the group. Also, in the case $p \gg n$, Lasso algorithms are limited because at most n variables can be selected. The ENET uses a mixture of the ℓ_1 (lasso) and ℓ_2 (ridge regression) penalties and can be formulated as:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{Enet}) = \left(1 + \frac{\lambda_2}{n}\right)\left\{\underset{\beta}{argmin}\left|\left|\boldsymbol{y} - \boldsymbol{X\beta}\right|\right|_2^2 + \lambda_2\left|\left|\boldsymbol{\beta}\right|\right|_2^2 + \lambda_1\left|\left|\boldsymbol{\beta}\right|\right|_1\right\}$$

Where, λ_2 is the Ridge penalty parameter, penalizing the sum of the squared regression coefficients and λ_1 the Lasso penalty, penalizing the sum of the absolute values of the regression coefficients. The Elastic net is used to perform both the ridge and the lasso regression by eliminating the ridge shrinkage effect while still keeping the de-correlation part of ridge regression, which is responsible for the grouping effect and then rely on the lasso shrinkage to achieve good prediction performance and sparsity (Zou and Hastie, 2003).

For the Elastic Net the regression coefficients are estimated as:

$$\beta^{e-net} = (\boldsymbol{X}'\boldsymbol{X} + \lambda_2\boldsymbol{I})^{-1}(\boldsymbol{X}'y - \frac{\lambda_1}{2}w)$$

Similar to the lasso, the elastic net simultaneously does automatic variable selection and continuous shrinkage, and is able to select groups of correlated variables. Elastic net is implemented in R using the package *elasticnet*.

3.5. Choice of Tuning Parameters

To select the optimal value of the tuning parameters we need disciplined way of selecting λ, i.e. we need to "tune" the value of λ. In their original paper, Hoerl and Kennard (1970) introduced ridge traces (plot of the components $of\ \beta_\lambda^{ridge}$ against λ) and suggested to choose λ for which the coefficients are not rapidly changing and have "sensible" signs. This technique however has no objective basis and is heavily criticized by many scholars. The standard practice now is to use

cross-validation technique. This method directly estimates the extra sample error, and chooses λ that minimizes the mean squared error.

Most commonly used cross validation approach is K-fold cross validation and below is the steps how it works.

(i) Partition the training data T into K separate sets of equal size.
Suppose $T = (T_1, T_2, \ldots, T_K)$. Commonly chosen K's are K = 5 and K = 10

(ii) For each $k = 1,2,\ldots,K$, fit the model $\hat{f}_{-k}^{(\lambda)}(\boldsymbol{X})$ to the training set excluding the k^{th} fold T_k

(iii) Compute the fitted values for the observations in T_k, based on the training data that excluded this fold

(iv) Compute the cross-validation (CV) error for the k^{th} fold:

$$(CV\ Error)_{\lambda}^{(k)} = |Tk|^{-1} \sum_{(X,y)\in T_k} \left(y - \hat{f}_{-k}^{(\lambda)}(\boldsymbol{X})\right)^2$$

The model then has overall cross-validation error:

$$(CV\ Error)^{\lambda} = K^{-1} \sum_{k=1}^{K} (CV\ Error)_k^{\lambda}$$

Finally, Select λ^* as the one with minimum$(CV\ Error)^{(\lambda)}$. Then refit the model with λ^* on the entire training set

One set of the cross validation is called leave-one-out cross validation, which has low bias but can have high variance. *Generalized cross-validation* provides a convenient approximation to leave one out cross-validation, for linear fitting under squared-error. (See Hastie, Tibshirani and Friedman, 2009; page ***214-215*** for more discussion on CV).

In this study, Cross-validation methods (both K-fold CV and GCV) were used to select the tuning parameter in ridge, lasso and elastic net models.

4. Results

4.1. Exploratory Data Analysis (EDA)

Table 1 below shows summary statistics of the Biomarker of effect (TSH) and some Exposure of biomarkers. It can clearly be seen that some of the continuous variables tend to have high nominal values and high standard errors than others, possibly due to the difference in measurement units used. For further statistical analysis in lasso, ridge regression and elastic net, we centered the response (by substracting $\bar{y}$ from y_i) and standardized the covariates (by setting unit variance) to take in to account the difference in measurement units.

Table 1: Summary statistics of variables.

	TSH	TRA	DL_PCB	MBZP	DCP	HHCB	TTMA	CB107	CB146
Min.	0.35	0.15	0.02	3.4	0.2	301	11	2	2
Mean	2.30	6.09	0.12	52.74	3.54	763.80	151.15	11.82	11.51
Std.D	0.95	3.89	0.05	70.47	7.21	272.33	217.88	8.10	7.66
Max.	6.43	24.2	0.29	706	65	1539	1649	77	77

	CB187	UCU	UAS	UCD	UTL	PCB118	PCB170	SOMPCB	COT	UBPA
Min.	2	3.85	2.7	0.05	0.06	20	10	44	10	0.10
Mean	15.19	11.29	24.4	0.28	0.22	30.98	29.89	262.10	623.03	3.35
S.E.	6.77	6.52	54.17	0.18	0.11	21.57	33.48	238.87	1584.98	5.37
Max.	57	47.83	520.4	1.11	0.70	170	350	2881	8156	53.40

	BDE47	BDE153	DDE	HCB	PCDD_F	MBP	HPYR	SOMME	UTCS
Min	5	1	51	10	0.11	5	24	8.3	0.10
Mean	6.18	3.16	415.72	40.18	0.40	55.90	176.44	83.59	26.27
S.E	3.69	2.95	444.57	16.99	0.13	107.01	196.02	148.87	82.15
Max.	26	23	3219	94	0.86	1420	2354	1895	705.97

Also, table A1 in the appendix gives the various catergories of categorical variable, presented as frequencies and expressed in percentages. The Categorical variables are gender, season, education level of the adolescent, body mass index (BMI), creatinine level, blood fat, sickness in the last two weeks and highest education level in the family.

4.2. Exposure Assessment: Impact of the Covariates on the Response

We now look at the relationship between responses with covariates using scatter plots. Figure 2, below shows a scatter plot for biomarker of effect (Log(TSH)), and few selected biomarkers of exposure(PCB170, PCB118, HCB and DDE). We generally observe from these four plots that, existence of EDCs in human body have a biomarker effect indicated by positive value of TSH

and it seems as with increase in EDCs exposure there is a moderate rising effect on TSH. Vertical trend observed on the scatter plot of PCB170 and HCB is as a result of imputation of one single value (LOD/2) for concentrations that are non-detectable or near limit of detection (LOD).

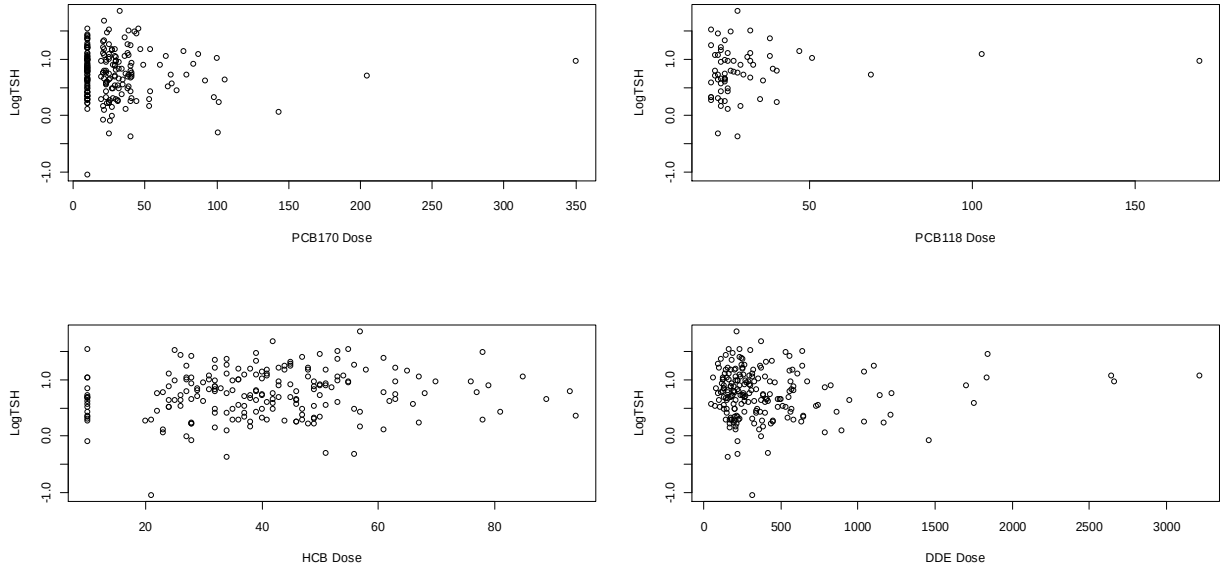

Figure 2: Dose-response plot between selected EDCs with Thyroid Stimulated Hormones.

Figure 3, below, presents scatter plot for persistent brominated pollutants (BED153 and BDE47) with thyroid stimulated hormones in log scale. There is a moderate relationship between brominated pollutants with thyroid stimulated hormones. The scatter plots for other covariates presented in figure A on appendix section. Lesser observations in PCB 118 (in figure 2) and BDE 47 (in figure 3) and on other variables presented in scatter plot figures in appendix is as a result of missingness. For missing values single imputation has done in our analysis in regularization methods.

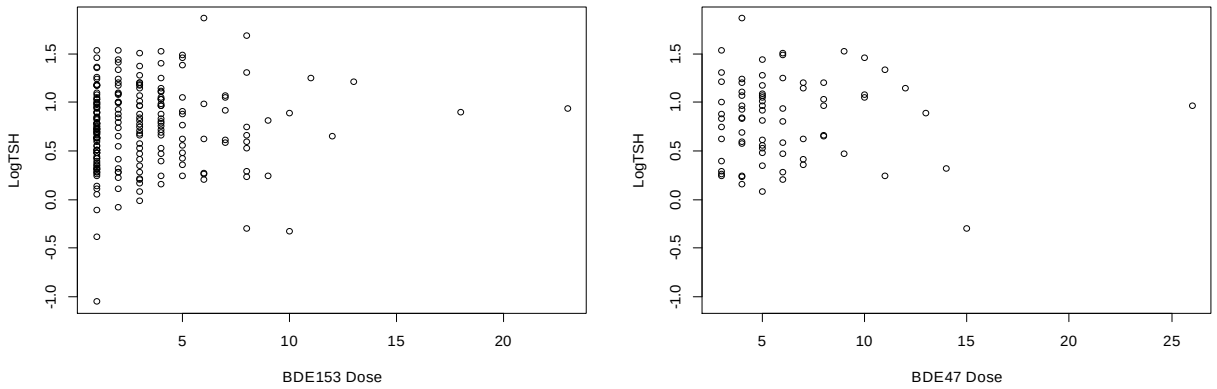

Figure 3: Dose-response plot between Brominated Pollutants (BED) with Thyroid Stimulated Hormones.

4.3. Multicollinearity Problem

In regression models problems arise when a serious multicollinearity (correlation among predictors) is present in the data, especially when there are numerous covariates. When the regressors are nearly perfectly related, the regression coefficients tend to be unstable, large standard error and the inferences based on the regression model can be misleading and erroneous (Gunst and Webster, 1975).

The most widely used techniques to indicate the presence of multicollinearity are large values for correlation between pairs of predictor variables in the correlation matrix, rxx (Neter, Wasserman and Kutner, 2005) and high Variance Inflation Factor (VIF) usually greater than 10. In this study, the associations/correlations among the predictors/effect modifiers are studied using pair-wise correlation coefficient matrix and VIF and the results were presented in Table A2 and Table A3 respectively in the appendix. High correlations are obtained among PCB170, SOMPCB, CB146, CB107 and PCDD_F. Enormous techniques are developed to handle problems due to multicollinearity (Agresti 2002, Neter, Wasserman, and Kutner 2005). Our objective in this study, however, is to apply shrinkage methods like Ridge, Lasso and Elastic Net techniques to handle the problem of multiculinarity.

4.4. Ridge Regression

In this study, we performed ridge regression and compared the result with OLS method depending on the standard deviation for individual estimators. To perform the ridge regression, we first standardized the covariates and centered the response in adjustment to differences in measurement units. Then we divide the Biomnitoring data in two parts: Test set (1/3 of the data) and Training set (2/3 of the data) to carry out model fitting and tuning parameter selection. To select the tuning parameter, generalized cross-validation (GCV) technique were carried out on the training data. The chosen λ_2 value is the one giving the smallest GCV error ($\lambda_2 = 1\ in\ our\ case$). Then we use this value of lambda to fit the entire model.

Ridge regression and OLS Parameter estimates with standard error was presented in appendix Table A3. From the results we can see that the standard error for each covariate in the Ridge regression estimate is less than that of OLS estimate. Consequently, we can conclude that Ridge regression model is better than OLS when the multicollinearity problem exists in the data. However, a ridge regression cannot produce parsimonious model, for it always keeps all the predictors in the model.

Figure 4, below, shows the estimated coefficients using ridge regression with various penalty parameters. For $\lambda_2 \in [0,100]$, the estimated coefficients approaches to zero as value of λ_2 increases.

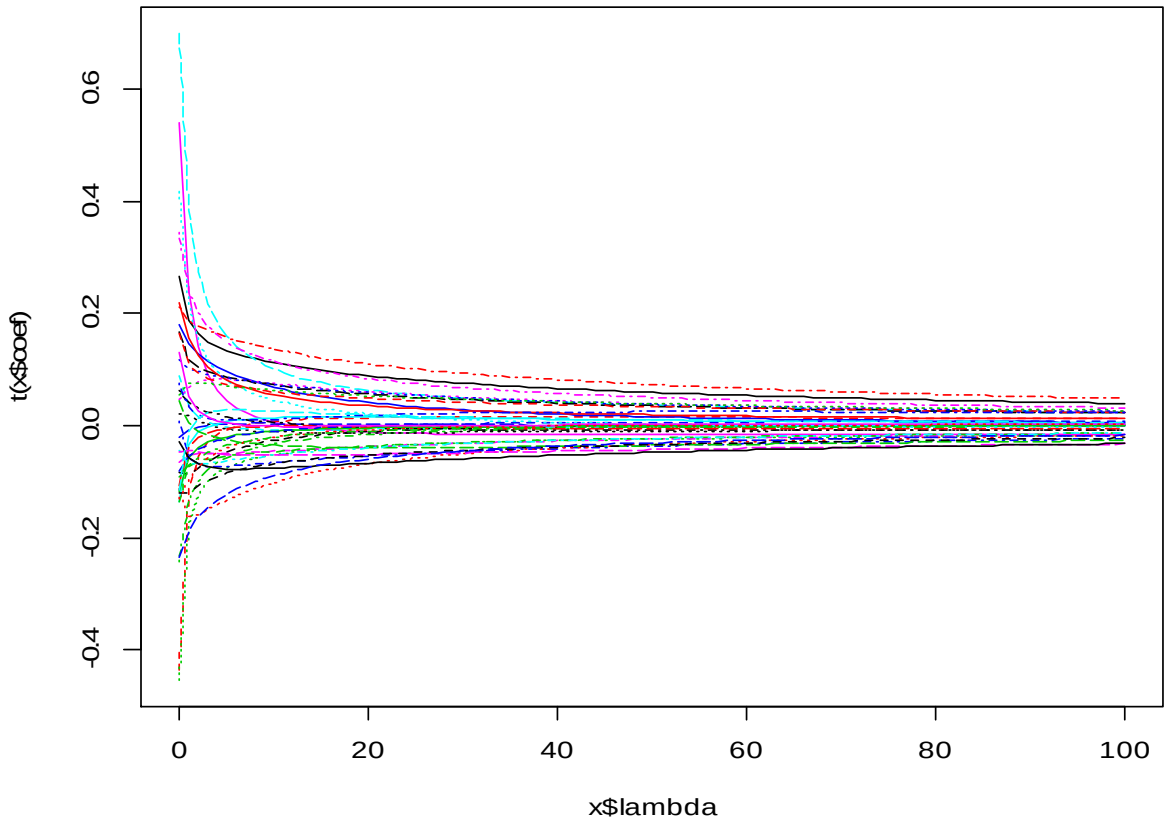

Figure 4: Estimated coefficients using ridge regression with various penalty parameters (X-axis = λ and Y-axis= $coefficients$).

4.5. The LASSO

Likewise in Ridge regression, before fitting the lasso model, we standardized the covariates and centered the response in adjustment to differences in measurement units. Then, we evaluated the lasso model on a grid of 20 *s* (relative bound) values. The trend of lasso parameter s (or tuning parameter λ_1) by Generalized Cross validation (GCV) scores can be observed on figure 5. From the figure, one can see that the speed of GCV scores decreases dramatically until *s* =0.6. And this value can be chosen as the optimal value of the tuning parameters as coefficients stabilizes (are not rapidly changing at this value). We have used the one-standard-error rule to pick the most parsimonious model within one standard error of the minimum.

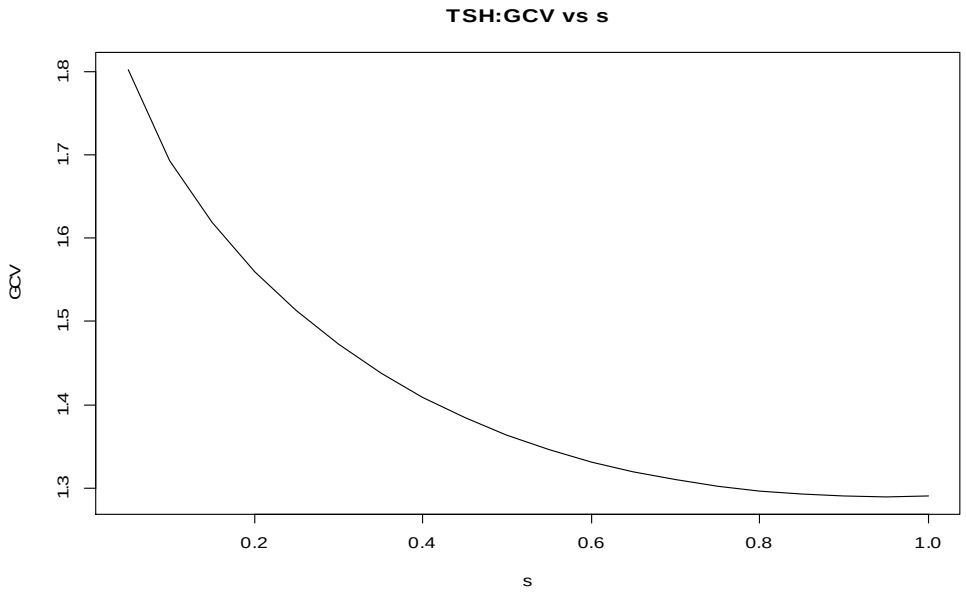

Figure 5: GCV score as a function of relative bound s. s = 0.6

In order to see how lasso shrinks and predicts the coefficients more clearly, the lasso estimates as a function of the standardized tuning parameter *s* as plotted. Intuitively, every coefficient will be squeezed to zero as *s* goes to zero.

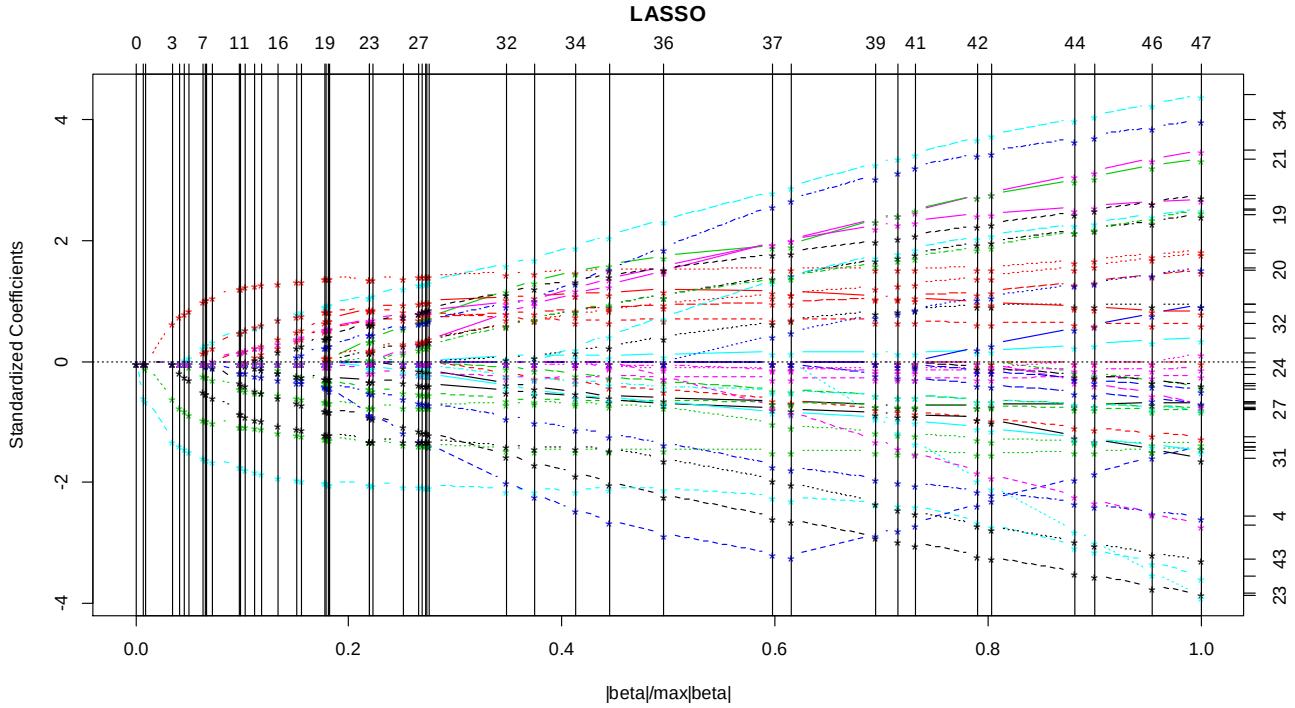

Figure 6: LASSO coefficient shrinkage in TSH data: each monotone decreasing curve represents a coefficient as a function of relative bound s. The vertical lines indicate the transition points; at the top the number of active predictors is given. The covariates enter the regression equation sequentially as s increase.
The predictors are (23) SOMPCB, (43) SEIZOEN, (4) UAS, (31) RB_A03, (27) UTL, (24) UBPA, (20) HCB, (19) DDE, (21) PCB118.

Finally, we fitted log (TSH) data at s = 0.6. To compute the standard error of the covariates, bootstrap method is used as suggested by Tibshirani (1996). The standard errors (Std.Error) were estimated by bootstrap re-sampling of residuals from the original data set by fixing s at its optimal value 0.6. Table A4 on appendix presents parameter estimates with their standard error (SE) of lasso regression. Our lasso model selected all covariates except SOMME, CB187, SOMPCB, SEZION2, Highest education level (lower secondary), BMI (25-30) and UCU which have zero coefficients. The possible reason for exclusion of these covariates is lasso drops one of the correlated variables and keeps the other and does not care which one is selected. For example the correlation between CB187 with CB146 (ρ=0.6) and PCB170 with SOMPCB (ρ=0.9), the model keep CB146 and PCB170 but dropped the others. We also noticed that lasso yielded smaller SE estimates than those of OLS. This is due to its constraint nature, that all predictions are subtracted by a threshold value.

4.6. The Elastic Net

Elastic net method, is an automatic variable selection method, naturally overcomes the difficulty of $p > n$ and has the ability to do grouped selection. Model fitting and tuning parameter selection by 10-fold cross validation were carried out on training data as described in Zou, H. & Hastie, T. (2003). There are two tuning parameters in the elastic net, so we need to cross-validate on a 2-dimensional surface. λ values (0, 0.01, 0.1, 1, 10 and 100) and s values (0, 0.1, 0.2, 0.3, 0.4, 0, 0.6, 0.7, 0.8, 0.9, and 1) were used to calculate cross validation error. The values of λ and s with the smallest cross validation error were considered for subsequent elastic net analysis. These were λ = 0.01 and s =0.9 as shown in Table 2 below. Parameter estimate with standard error using Elastic net model is presented on Table A5 in appendix section. This technique has an advantage of both model selection and inclusion more correlated covariates compared to Lasso.

Table 2: Cross validation errors for elastic net tuning parameters.

λ	***s=0***	***s=0.1***	***s=0.2***	***s=0.3***	***s=0.4***	***s=0.5***	***s=0.6***	***s=0.7***	***s=0.8***	***s=0.9***	***s=1***
0	0.168	0.177	0.177	0.169	0.156	0.148	0.143	0.141	0.139	0.138	0.144
0.01	0.150	0.163	0.172	0.171	0.167	0.160	0.155	0.148	0.142	**0.138**	0.139
0.1	0.142	0.161	0.171	0.178	0.182	0.186	0.191	0.200	0.210	0.216	0.221
1	0.160	0.176	0.182	0.176	0.173	0.175	0.185	0.199	0.209	0.228	0.246
10	0.199	1.381	3.114	4.801	5.887	6.610	7.494	8.326	9.164	10.069	10.938

A plot of the standardize coefficients of the parameters from the elastic net as a function of s are displayed on Figure 4.

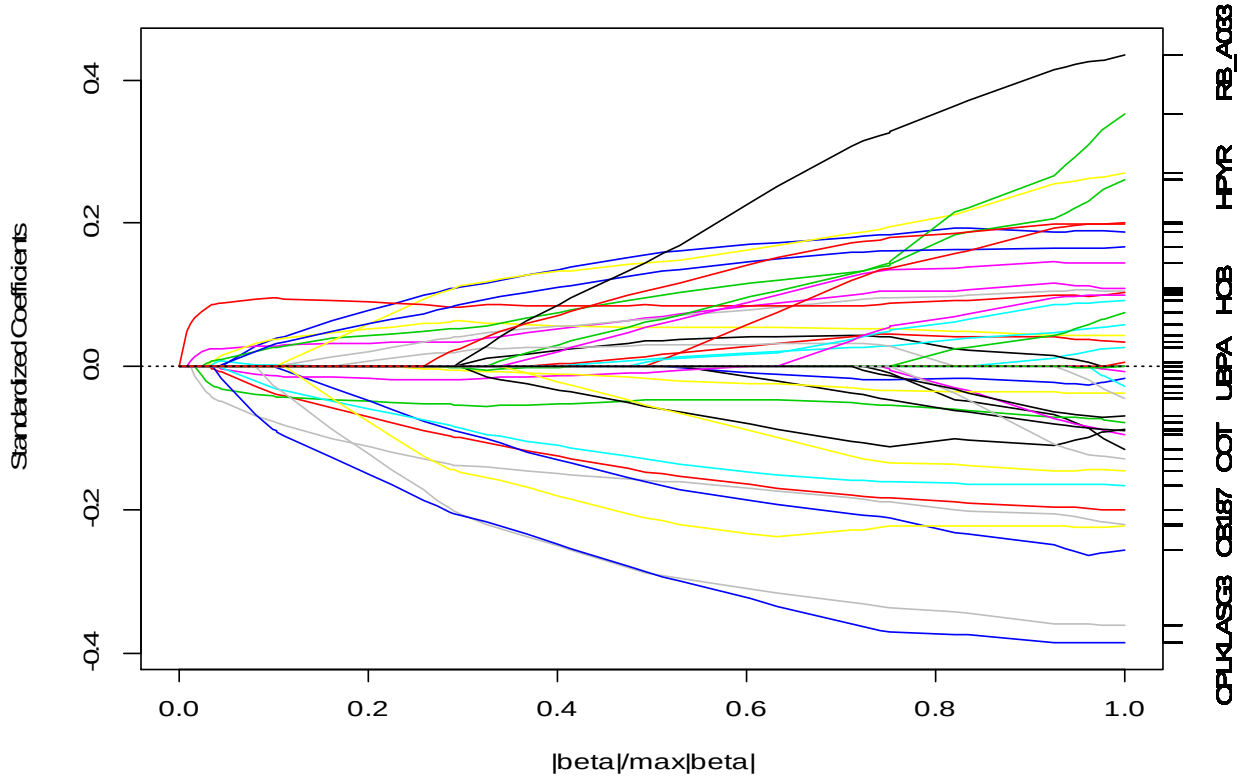

Figure 4: Standardize elastic net coefficients as a function of s =|beta|/max|beta|

4.7. Comparing Different Methods

We compared the performance of OLS, ridge, lasso and elastic net by computing their mean prediction mean-squared error on the test data. Table 3 clearly shows ridge, lasso and elastic net perform better than that of OLS having lower MSE and MSPE. The elastic net is the best option among the methods considered in terms of prediction accuracy because of it has smallest mean square error and mean square prediction error. It is clear that all regularization methods improve the prediction accuracy compared to no shrinking (OLS).

Table 3: TSH data: Comparing different methods

Method	**Parameter**	**Mean-squared error**	**Mean Squared Prediction Error**
OLS		0.923	1.913
Ridge	$\lambda = 1$	0.912	1.563
LASSO	s=0.6	0.881	1.361
Elastic net	λ =0.01, s=0.9	0.733	1.329

5. Conclusion and Discussion

This paper as a part of the project entitled “Human Biomonitoring: Multiculinarity” and aimed to investigate the influence of a number of covariates (levels of pollutants, gender, life style...) on Thyroid Stimulated Hormone. On this study we had 210 observations, with 35 covariates. 14 observations were deleted due to existence of missing values in categorical covariates, however single imputation technique is used to fill in missing values of continuous covariates. Standardizing is done to take in account the large difference in values of continuous covariates, which is due to measurement unit difference.

We detected for multicollinearity problem, using pair-wise correlation and VIF methods. Using these methods high correlations are obtained among PCB170, SOMPCB, CB146, CB107 and PCDD_F covariates. Multiple regression is a technique to investigate the influence of a number of covariates on the response but this methods fails because of inclusion of highly correlated covariates in the model. To take into account these limitation valid techniques such as: Ridge, Lasso and Elastic Net regression were suggested, and we investigate if these methods are valuable in our case.

In all the three models to validate the result one third observations was sampled without replacement at random as test set while the remaining two third observations was used as the training set. To choose the optimal value of tuning parameter we used GCV and 10-fold cross-validation techniques. We then compared the performance of those methods by computing their prediction mean squared error on the test data.

Ridge regression is a continuous shrinkage method that minimizes residual sum squares (RSS) subject to bound on the ℓ_2 norm of the coefficients. In Ridge regression, the coefficients are shrunken towards zero and will never become exactly zero unless $\lambda = \infty$. This introduces some bias, but can greatly reduce the variance, resulting in a better mean-squared error. However, when the number of predictors is large, Ridge regression will not provide a sparse model that is easy to interpret.

Using the generalized cross validation method we obtained optimum λ_2 equal to 1 and we fitted the ridge regression at this value. This regression gives estimates for all covariates with a lower

standard of error compared to OLS. However, ridge regression cannot produce a parsimonious model, for it always keeps all the predictors in the model.

Among methods that do both continuous shrinkage and variable selection, a promising technique called the *Least Absolute Shrinkage and Selection Operator* (lasso) was proposed by Tibshirani (1996). It is a penalized least square method imposing an ℓ_1 penalty on the regression coefficient. It reduces the variability of the estimates by shrinking the coefficients and at the same time produces interpretable models by shrinking some coefficients to exactly zero.

We fit a lasso model for log (TSH) on a grid of 20 *s* values. The optimal lasso parameter *s* was selected by generalized cross validation scores. We obtained 0.6 as our optimized *s* value, and fitted Lasso at this value. Our Lasso model provided mean estimates for COT, UAS, TRA, DL_PCB, MBP, MBZP, DCP, HHCB, TTMA, HPYR, BDE47, BDE153, CB107, CB146, DDE, HCB, PCB118, PCB170, UTCS, UTL, sickness in last 2 weeks, blood fat and gender are non-zero coefficients. This, technique produced a relatively parsimonious model, with lower standard error as compared to OLS.

The elastic net penalty is a double shrinkage method and it provides a compromise between the lasso and ridge penalties. It has the effect of averaging markers that are highly correlated and then entering the averaged marker into the model. The values of λ and *s* that gives the smallest cross validation error were considered for elastic net analysis. We selected $\lambda = 0.01$ and $s = 0.9$ as optimal tuning values. We fitted elastic net regression at these values and obtained parameter estimates for all covariates except SOMPCB, UBPA and UCD.

Finally, we compared the performance of those methods by computing their prediction mean-squared error on the test data. It is clear that all regularization methods improve the prediction accuracy compared to no shrinking (OLS). The Elastic Net performs better than the Lasso, and the Lasso performs better than Ridge regression.

In conclusion, in biomonitoring studies which involve several covariates and problem of multicolinearity, Ridge, Lasso and Elastic net regression techniques are valuable approaches.

6. Reference

Agresti, A. (2002). *Categorical Data Analysis (*2nd ed.). New York: John Wiley & Sons.

Brucker-Davis, F. (1998). Effects of environmental synthetic chemicals on thyroid function. *Thyroid*, **8**, 827–856.

Diamanti-Kandarakis, E. et al. (2009). Endocrine disrupting chemicals: an Endocrine Society scientific statement. Endocrine Reviews, 30(4):293–342.

EFSA (2005). Opinion of the scientific panel on contaminants in the food chain on a request from the commission related to the presence of non dioxin-like polychlorinated biphenyls (PCB) in feed and food. *The EFSA Journal*, **420**, 1 – 34.

EFSA (2006a). Opinion of the scientific panel on contaminants in the food chain on a request from the commission related to DDT as an undesirable substance in animal feed. The EFSA Journal, 433, 1 – 69.

EFSA (2006b). Opinion of the scientific panel on contaminants in the food chain on a request from the commission related to hexachlorobenzene as an undesirable substance in animal feed. *The EFSA Journal*, **402**, 1 – 49.

Friedman, J., Hastie, T. and Tibshirani R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical software, **33**:1-22.

Gunst, R.F., and Webster, J.T. (1975). Regression analysis and problem of multicollinearity. Communications in Statistics-theory and Methods, 4, 277-292.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning (2ndEd.). Springer, New York, USA.

Hites, R.A. (2004). Polybrominated diphenyl ethers in the environment and in people: a meta-analysis of concentrations. *Environ Sci Technol*. **38**(4):945–956.

Hoerl, A. and Kennard, R. (1988). Ridge regression, in 'Encyclopedia of Statistical Sciences', Vol. 8, Wiley, New York, pp. 129–136.

Hoerl, A.E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**: 55-67.

Hoh, E. and Hites, R.A. (2005). Brominated flame retardants in the atmosphere of the east-central United States.Environ Sci Technol.**39**(20):7794–7802.

Howdeshell, K.L. (2002). A model of the development of the brain as a construct of the thyroid system. *Environmental Health Perspectives Supplements*, **110**, 337–348.

Neter, J., Wasserman, W., and Kutner, M.H. (2005). Applied Linear regression Models (5th ed.). McGraw Hill/Irwin Series: *Operations and Decision Sciences*.

Scientific Co-operation on Questions Relating to Food (SCQRF) (2000). Assessment of dietary intake of dioxins and related PCBs by the population of EU member states. Directorate-General Health and Consumption Protection.

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

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society B*, **67**: 301-320.

Zou, H. and Hastie, T. (2003). Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays. *Department of Statistics, Stanford University*.

7. Appendix

A. Tables and Figures

Table A1: Summary statistics of categorical variables.

Variable	Category	Freq	Variable	Category	Freq	Variable	Category	Freq
Gender	Male	121(57.6%)	**Sickness in Last two weeks**	Yes	165(80.5%)	**SEIZON**	winter	62(29.5%)
	Female	89(42.4%		No	45(21.4%)		Spring	101(48.1%)
							Autumn	47(22.4%)
Variable	**Category**	**Freq**	**Variable**	**Category**	**Freq**	**Variable**	**Category**	**Freq**
Blood Fat	<400mg/week	55(26.2%)	**Creatinine level**	<95mg/dl	55(26.2%)	**Educat. Level of the adolescent**	General	101(48.8%)
	[400-430]	51(24.3%)		[95, 135]	51(24.3%)		Technical	86(41.6%)
	[430-480]	51(24.3%)		[135, 190]	53(25.2%		Vocational	23(11%)
	>=480	53(25.2%)		>=190	51(24.3%)			
Variable	**Category**	**Freq**	**Variable**	**Category**	**Freq**			
BMI	[18.5-25] kg/m2	20(9.5%)	**Highest Education Level in the family**	Max. lower secondary	28(13.3%)			
	[25-30]	168(80%)		Max. High secondary	66(31.4%)			
	>=30	22(10.5%)		Higher education level	116(55.2%)			

Table A2: pair-wise Correlation coefficients for the covariates

	TSH	DL_PCB	MBZP	SOMME	HHCB	CB107	CB146	CB187	COT	UCU	UAS	PCB118	PCB170	UBPA	SOMPCB	PCDD_F	MBP	DCP	HPYR	BDE47	BDE153
TSH	1.00																				
DL_PCB	0.03	1.00																			
MBZP	0.12	0.09	1.00																		
SOMME	-0.04	0.08	0.12	1.00																	
HHCB	0.03	-0.04	-0.05	-0.05	1.00																
CB107	0.02	-0.02	0.13	0.08	-0.03	1.00															
CB146	-0.04	0.08	0.12	0.06	0.08	0.68	1.00														
CB187	-0.07	0.21	0.03	-0.04	0.21	0.36	0.61	1.00													
COT	-0.05	-0.34	0.15	0.00	-0.14	-0.02	-0.05	-0.15	1.00												
UCU	0.04	0.00	0.27	0.15	0.00	0.20	0.12	-0.08	0.33	1.00											
UAS	-0.03	0.12	0.06	0.04	-0.09	0.01	-0.01	-0.16	-0.07	0.25	1.00										
PCB118	0.09	-0.20	0.09	0.03	-0.09	0.51	-0.01	0.24	0.18	-0.03	-0.07	1.00									
PCB170	-0.06	0.10	-0.01	0.00	0.00	0.19	0.51	0.47	-0.09	-0.09	0.02	0.68	1.00								
UBPA	-0.03	-0.09	0.03	0.04	-0.06	0.01	0.05	0.01	0.04	0.20	-0.03	0.01	0.00	1.00							
SOMPCB	-0.03	0.08	0.07	0.02	-0.11	0.32	0.56	0.38	-0.04	-0.06	0.04	0.78	0.92	0.00	1.00						
PCDD_F	0.03	0.64	0.01	-0.06	-0.10	0.14	0.09	0.13	-0.33	0.00	0.04	-0.27	0.09	-0.07	0.07	1.00					
MBP	0.10	-0.01	0.23	0.05	-0.08	0.16	0.07	0.01	-0.01	0.34	-0.01	-0.04	-0.01	0.06	0.01	0.13	1.00				
DCP	0.19	-0.11	0.11	-0.01	-0.07	-0.08	-0.05	-0.04	-0.05	0.19	0.09	-0.05	-0.07	0.03	-0.04	-0.14	-0.01	1.00			
HPYR	0.08	-0.12	0.21	0.03	0.14	0.02	0.00	-0.02	0.55	0.56	0.00	0.04	-0.07	0.07	-0.07	-0.14	0.16	0.03	1.00		
BDE47	-0.01	0.08	0.02	0.17	-0.20	0.08	0.16	-0.04	0.26	0.07	0.00	-0.19	-0.02	0.11	0.02	0.33	0.02	-0.01	-0.03	1.00	
BDE153	0.07	-0.06	0.09	0.03	-0.21	0.27	0.24	0.03	0.06	0.07	0.00	-0.02	0.05	0.01	0.09	0.00	0.03	0.07	-0.02	0.00	1.00

Table A3: Ridge and OLS estimate and Std,Error

	Ridge regression			OLS			
Label							VIF
	Estimate	Std.error	P- value	Estimate	Std.error	P-value	
Intercept	0.690	0.807	0.398	0.181	0.231	0.445	0
Gender	0.157	0.154	0.315	0.054	0.335	0.874	3.193
COT	-0.177	0.201	0.276	0.261	0.181	0.169	5.519
TRA	-0.026	0.023	0.789	0.339	0.242	0.181	10.857
DL_PCB	0.513	1.906	0.909	0.676	0.432	0.138	3.951
MBZP	0.000*	0.002	0.531	-0.151	0.270	0.584	2.850
SOMME	0.000*	0.001	0.454	0.757	0.911	0.419	2.360
HHCB	0.000*	0.000	0.242	-0.281	0.284	0.338	2.112
CB107	0.019	0.016	0.492	0.549	0.524	0.311	16.103
CB146	-0.015	0.022	0.943	-0.589	0.610	0.349	18.799
CB187	-0.001	0.017	0.020	0.484	0.372	0.213	6.871
UCD	1.526	0.628	0.414	-0.054	0.228	0.816	3.739
UCU	0.014	0.017	0.211	-0.010	0.530	0.985	3.049
UTL	-0.802	0.631	0.028	0.000	0.180	0.999	2.550
UAS	-0.003	0.001	0.886	-0.095	0.556	0.867	2.940
PCB118	-0.246	0.193	0.784	0.178	0.590	0.767	7.91
PCB170	0.001	0.004	0.356	0.061	0.863	0.945	14.468
UBPA	-0.002	0.007	0.656	-0.014	0.150	0.927	1.765
UTCS	-0.001	0.001	0.083	-0.136	0.165	0.425	2.555
SOMPCB	0.000*	0.001	0.009	0.290	1.103	0.796	25.687
High Educ.lev 1	-0.331	0.186	0.329	0.192	0.406	0.643	4.088
High Educ.lev 2	-0.458	0.167	0.088	-0.001	0.442	0.999	3.858
BMI(18.5-25)	-0.263	0.266	0.304	-0.457	0.338	0.197	4.766
BMI(25-30)	-0.358	0.204	0.267	-0.495	0.317	0.139	4.277
HCB	0.004	0.004	0.239	-0.081	0.247	0.748	2.421
DDE	0.000*	0.000	0.313	-0.233	0.240	0.346	3.295
PCDD_F	-0.646	0.540	0.448	0.009	0.342	0.979	3.094
MBP	-0.003	0.003	0.558	-0.046	0.149	0.763	2.985
DCP	0.015	0.019	0.054	0.179	0.273	0.522	2.799
HPYR	0.000*	0.001	0.097	0.323	0.155	0.055	3.629
TTMA	0.001	0.000	0.063	-0.055	0.172	0.753	2.043
BDE47	-0.001	0.121	0.313	-0.280	0.158	0.097	3.022
BDE153	0.028	0.016	0.136	0.414	0.193	0.049	1.446
Sickness	-0.318	0.166	0.946	-0.260	0.243	0.300	1.704
Blood fat(<400)	0.198	0.194	0.438	-0.340	0.248	0.190	3.989
Blood fat(400-430)	0.259	0.170	0.398	0.283	0.289	0.342	3.337
Blood fat(430-480)	0.013	0.185	0.315	-0.361	0.292	0.236	3.083
Creatinine							

level(<95mg/dl)	0.215	0.274	0.276	0.027	0.334	0.937	9.119
Creatinine level(95-135)	0.396	0.246	0.116	0.690	0.295	0.034	7.020
Creatinine level(135-190)	0.335	0.183	0.075	0.129	0.258	0.626	3.308
Winter	1.048	0.542	0.061	-0.362	0.351	0.318	4.227
Spring	0.711	0.524	0.183	-0.853	0.464	0.086	4.052
General	-0.655	0.355	0.073	-1.501	0.672	0.041	5.702
Technical	-0.747	0.328	0.029	-2.066	0.746	0.014	4.766

**small value*

Table A4: LASSO model Parameter estimates with S.E

Coefficients	Estimate	Std.error	P-value
(Intercept)	0.107	0.102	0.295
GESL	-0.205	0.120	0.087
Sickness in last 2 weeks	-0.097	0.095	0.305
COT	-1.009	0.455	0.026
UAS	-0.224	0.079	0.005
TRA	0.415	0.130	0.001
DL_PCB	0.143	0.140	0.305
PCDD_F	-0.255	0.128	0.046
MBP	-0.685	0.471	0.146
MBZP	0.003	0.224	0.990
DCP	0.024	0.074	0.752
HHCB	0.146	0.147	0.321
TTMA	-0.113	0.107	0.292
HPYR	0.249	0.215	0.248
BDE47	0.395	0.180	0.029
BDE153	0.090	0.117	0.439
CB107	0.042	0.244	0.864
CB146	-0.333	0.227	0.143
CB187	0.000	0.147	1.000
DDE	0.170	0.119	0.151
HCB	0.211	0.104	0.042
PCB118	0.003	0.161	0.986
PCB170	-0.179	0.222	0.420
SOMPCB	0.000	0.306	1.000
UBPA	0.121	0.224	0.588
UTCS	0.003	0.123	0.980
UCD	0.105	0.129	0.416
UTL	-0.061	0.118	0.605
UCU	0.000	0.125	1.000
SOMME	0.000	0.165	1.000

General	-0.369	0.202	0.067
Technical	-0.353	0.204	0.084
Highest Educ.lev(lower scond)	0.000	0.131	1.000
Highest Educ.lev(high scond)	-0.034	0.135	0.801
Creatinine(<95)	0.333	0.144	0.021
Creatinine [95-135]	0.212	0.126	0.092
Creatinine [135-190]	0.268	0.121	0.027
Body fat (<400mg/week)	0.234	0.106	0.028
Blood fat [400-430]	0.228	0.108	0.036
Blood fat [430-480]	0.158	0.103	0.125
BMI[18.5-25]	0.035	0.129	0.787
BMI[25-30]	0.000	0.141	1.000
Winter	0.013	0.104	0.902
Spring	0.000	0.075	1.000

Table A5: Parameter estimates with S.E using Elastic net model

Coefficients	Estimate	Std.Error	P-value
(Intercept)	0.124	0.091	0.173
GESL	-0.059	0.116	0.612
Sickness in last 2 weeks	-0.107	0.097	0.272
COT	-0.177	0.201	0.378
UAS	-0.190	0.099	0.055
TRA	0.374	0.129	0.004
DL_PCB	0.289	0.138	0.036
PCDD_F	-0.334	0.128	0.009
MBP	0.431	0.379	0.256
MBZP	-0.032	0.127	0.798
DCP	-0.053	0.095	0.581
HHCB	0.033	0.123	0.788
TTMA	-0.024	0.124	0.846
HPYR	0.145	0.241	0.549
BDE47	-0.001	0.121	0.996
BDE153	-0.076	0.135	0.573
CB107	0.078	0.193	0.687
CB146	-0.285	0.213	0.182
CB187	0.011	0.150	0.942
DDE	0.207	0.120	0.085
HCB	0.134	0.122	0.273
PCB118	0.246	0.193	0.203
PCB170	-0.109	0.274	0.692
SOMPCB	0.000	0.127	1.000
UBPA	0.000	0.121	1.000

UTCS	-0.039	0.122	0.749
UCD	0.000	0.119	1.000
UTL	-0.079	0.120	0.512
UCU	0.117	0.121	0.336
SOMME	-0.057	0.080	0.476
General	-0.059	0.191	0.757
Technical	-0.139	0.178	0.434
Highest edu. (lower secondary)	0.056	0.148	0.707
Highest edu.(high secondary)	-0.122	0.150	0.415
Creatinine (<95mg/dl)	0.349	0.165	0.034
Creatinine [95-135]	0.213	0.158	0.178
Creatinine [135-190]	0.234	0.141	0.096
Blood fat [<400 mg/week]	0.240	0.126	0.057
Blood fat[400-430]	0.151	0.121	0.211
Blood fat [430-480]	0.211	0.123	0.088
BMI[18.5-25]	-0.037	0.125	0.767
BMI[25-30]	-0.126	0.132	0.341
Winter	-0.227	0.161	0.160
Spring	-0.281	0.161	0.082

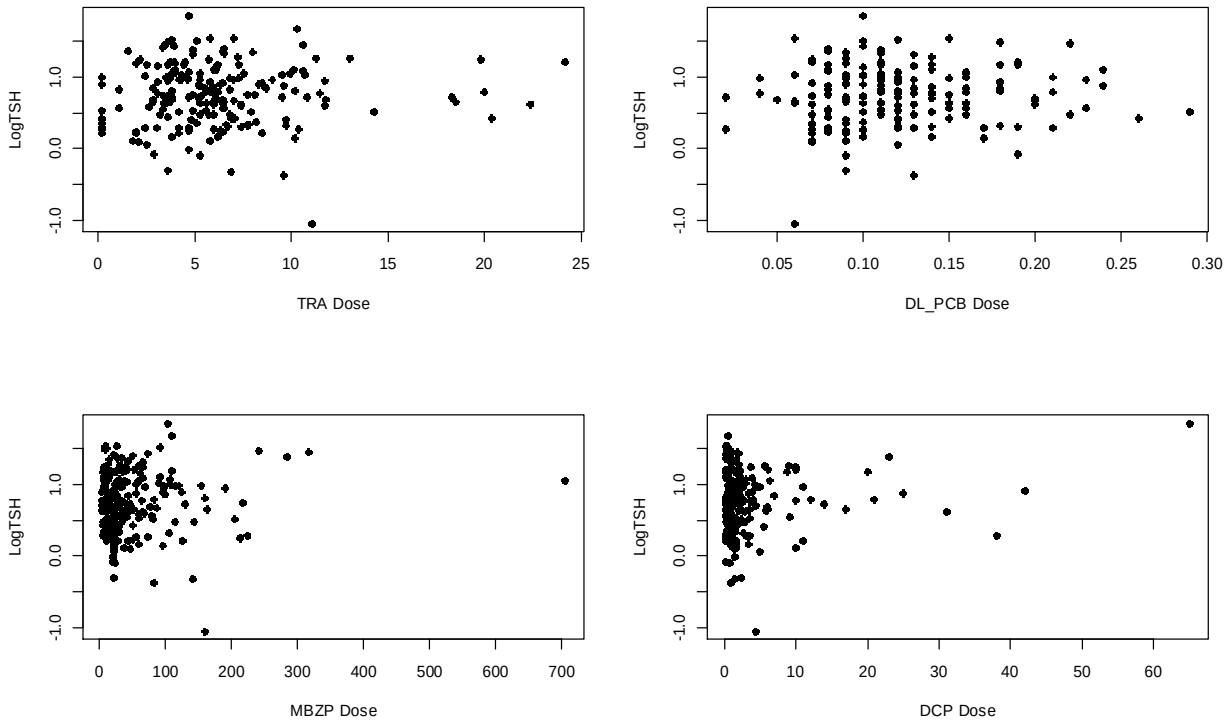

Figure A1: Dose-response plot between Pollutants (TRA, DL_PCB, MBZP & DCP) with TSH.

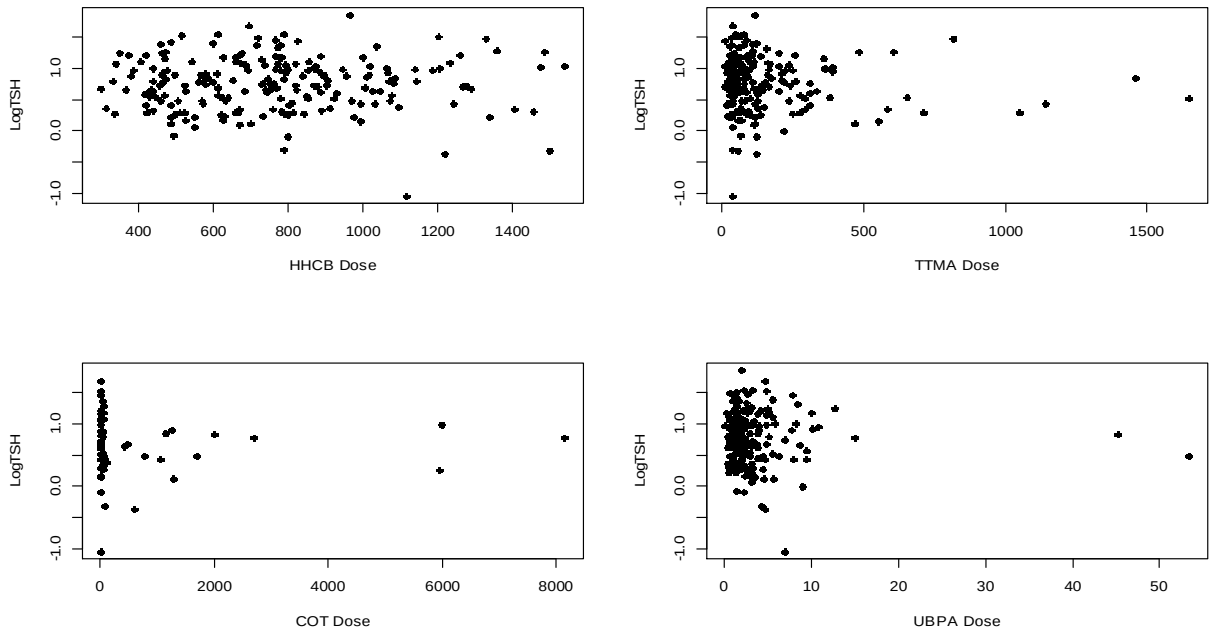

Figure A2: Dose-response plot between Pollutants (HHCB, TTMA, COT & UBPA) with TSH.

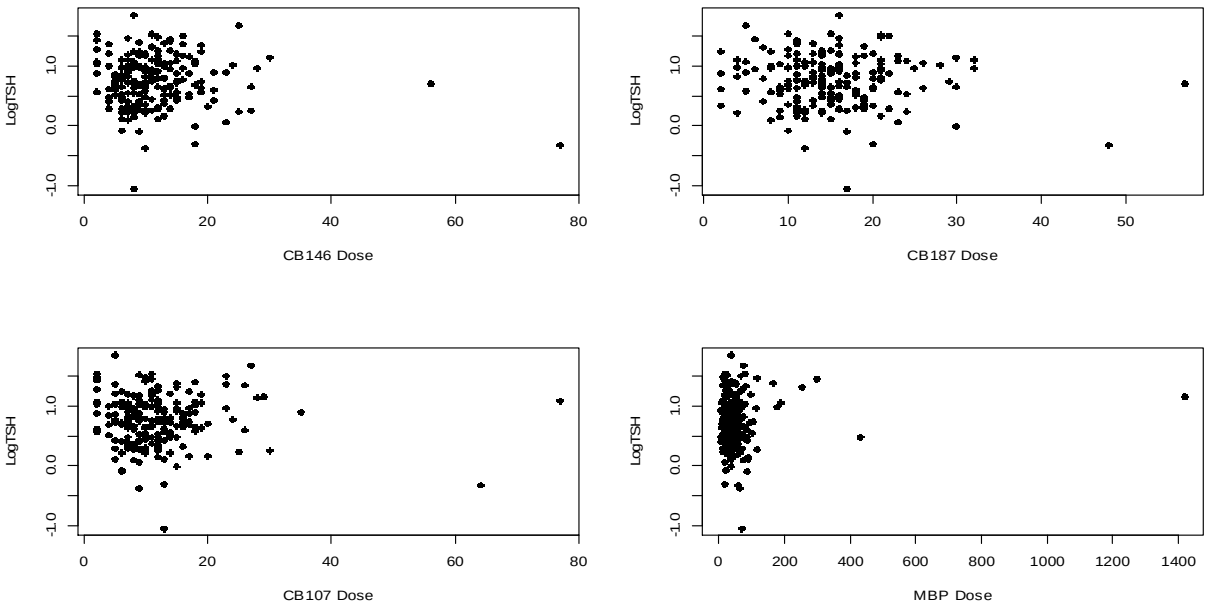

Figure A3: Dose-response plot between Pollutants (CB146, 187, 107 &MBP) with TSH.

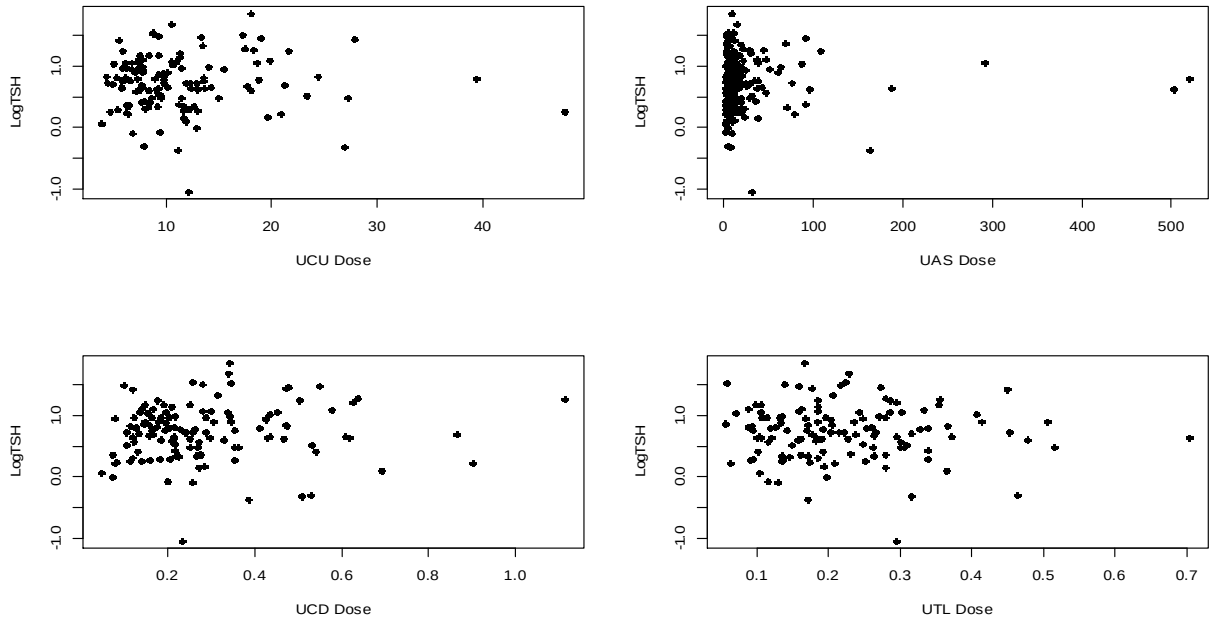

Figure A4: Dose-response plot between Pollutants (UCU, UAS, UCD & UTL) with TSH.

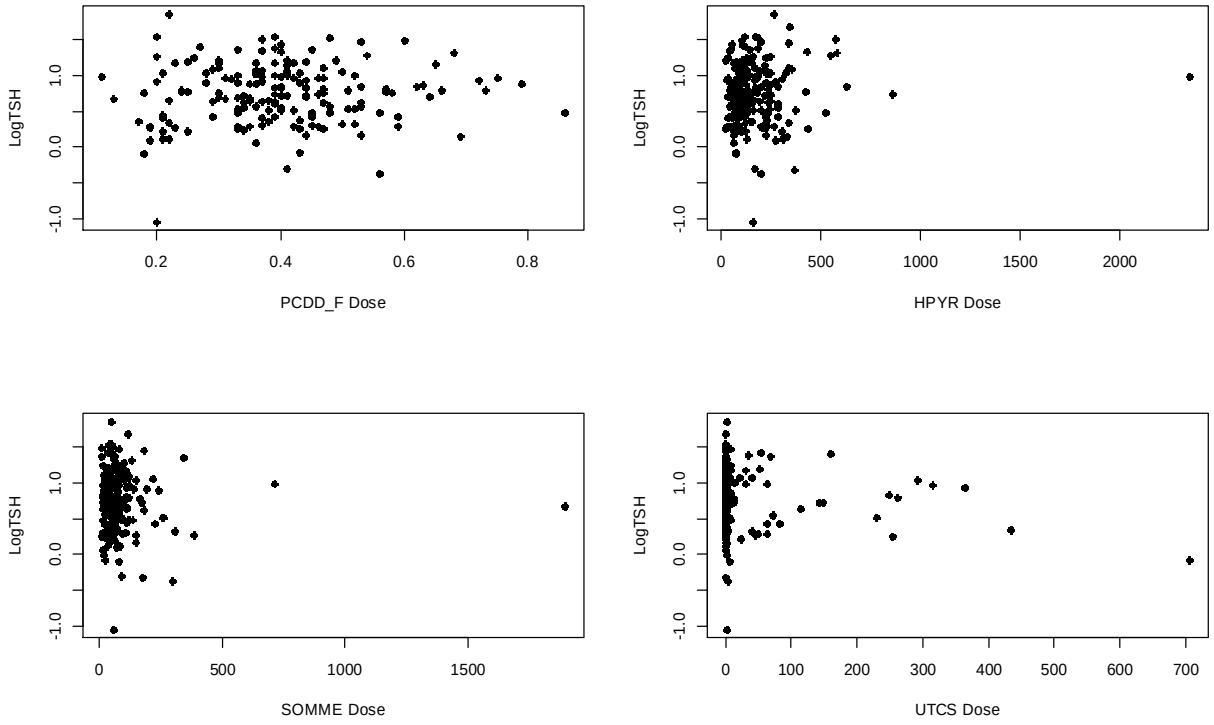

Figure A5: Dose-response plot between Pollutants (PCDD_F, HPYR, SOMME & UTCS) with TSH.

Selected R Code

```
library(lasso2)
library(elasticnet)
library(lars)
library(dummies)
library(MASS)
###### Trean and test sample######
set.seed(54321)
sample.size=length(dataridge[,1])
frac=0.7
learn.set=sample(sample.size,sample.size*frac)
learn.dat=dataridge[learn.set,]
test.dat=dataridge[-learn.set,]
library(MASS)
############## Ridge Regression###############
fit.ridge<lm.ridge(tsh1~GESL1+RB_A031+RB_A032+RB_F02+SEIZOEN1+SEIZOEN2+BMITOVO1+BMITOV
O2+DL_PCB+HHCB+SOMME+TRA+MBZP+DCP+DDE+TTMA+HPYR+UCD+UBPA+BLOEDVETKLAS1+B
LOEDVETKLAS2+BLOEDVETKLAS3+CREAKLAS1+CREAKLAS2+CREAKLAS3+SOMPCB+UAS+UCU+
UTL+UTCS+MBP+PCDD_F+BDE153+HCB+CB146+CB187+CB107+PCB170+HCB+OPLKLASG1+OPLKLA
SG2+PCDD_F
,standardize = TRUE, data=learn.dat,lambda=seq(0,1,0.01))
select(fit.ridge)
which.min(fit.ridge$GCV)

#####Cross validation procedure to decide tuning parameter t
cv.TSH<-cv.lars(X,Y,K=10,trace =F, plot.it = TRUE, se = TRUE,type = c("lasso"),mode=c("fraction"),
normalize = F)
cv.TSH
title("10-fold Cross Validation and Standard error")
lars_TSH<-lars(X,Y,type="lasso",intercept=TRUE)
plot(lars_TSH)
#Using "Lasso2" Package

l1c.tsh<- l1ce(tsh1 ~.,data=learn.dat,trace=TRUE, sweep.out=~1,bound=(1:20)/20)
l1c.tsh

###----Elastic Net----###
data.net<-as.matrix(data.dummies[-test,]);
xpred<-as.matrix(data.dummies[test,]);
s<-seq(0,1,0.1)
lambda<-c(0, 0.01, 0.1, 1, 10);
errors<-NULL;
##For lambda=c(0, 0.01, 0.1, 1, 10) and s=seq(0,1,0.1) get the combination with lowest CV error##
for(i in lambda){
        net.params<-cv.enet(data.net[,-28], data.net[,28], K=10, lambda=i, s=seq(0,1, 0.1),
mode="fraction",
        plot.it=FALSE, se=TRUE, max.steps=50, normalize=F);
        cve<-net.params$cv.error;
        errors<-rbind(errors,cve);
}
```

```
rownames(errors)<-lambda;
colnames(errors)<-s;
errors

enet.reg<-enet(data.net[,-28], data.net[,28], lambda=0.01, max.steps=50, normalize=T,intercept=T);
plot.enet(enet.reg, use.color = TRUE);
enet.reg
coef<-predict.enet(enet.reg, s=0.9, se=TRUE,type = "coefficients",intercept=T, mode = "fraction");
coef

pred.enet<-predict.enet(enet.reg, newx=data.net[,-28], s=0.9, type="fit", mode="fraction");
mse.learn<-mean((pred.enet$fit - data.net[,28])^2);
mse.learn;

pred.enetl<-predict.enet(enet.reg, newx=xpred[,-28], s=0.9, type = "fit", mode = "fraction");
mse.test<-mean((pred.enetl$fit - xpred[,28])^2)
mse.test
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Human biomonitoring: multicollinearity

Richting: **Master of Statistics-Epidemiology & Public Health Methodology**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Taye, Mekdes Abera

Datum: **24/09/2013**