Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

ABSTRACT

Comprehending the varied latent profiles among pregnant women in relation to toxicant exposures and levels of endogenous biomarkers is essential for guiding precise interventions and prenatal care strategies linked to gestation age. Moreover, such understanding may provide valuable insights into potential approaches for enhancing prenatal care and fostering healthy pregnancies. This study aims to discern distinct latent profiles among pregnant individuals based on their patterns of toxicant exposure and biomarker levels, and to examine how these profiles correlate with gestation age at the final visit. Leveraging a synthetic dataset comprising 161 observations, six covariates, and an outcome variable designed to mirror the relationship between gestational age at delivery and the phthalate risk score from the actual dataset "environment mediation framework" sourced from GitHub, latent profile analysis and factor analysis were employed. These analyses focused on toxicant exposures and endogenous biomarkers to delineate unique subgroups within the pregnant population. Additionally, we conducted a LASSO regression to assess how the identified unique profiles relate to gestation age at the final visit, while controlling for covariates. These findings will help illuminate patterns among pregnant women and their impact on gestational outcomes.

INTRODUCTION

Pregnancy is a critical period during which maternal exposures to various toxicants can potentially impact both maternal and fetal health outcomes. Understanding the complex interplay between toxicant exposure profiles, biomarker levels, and gestational outcomes is of paramount importance for optimizing prenatal care strategies and promoting healthy pregnancies. Despite the importance of this topic, there remains a gap in the literature regarding the combined effects of multiple toxicant exposures profiles and Endogenous biomarkers levels.

Background:

During pregnancy, maternal exposure to environmental toxicants, such as air pollutants, heavy metals, pesticides, and industrial chemicals, is of particular concern due to the potential adverse effects on fetal development (Grandjean & Landrigan, 2006). These toxicants can cross the placenta and accumulate in fetal tissues, posing risks for various adverse outcomes, including preterm birth, low birth weight, and developmental abnormalities (Braun et al., 2016; Wigle et al., 2008).

Furthermore, biomarkers play a crucial role in assessing internal exposure levels and understanding the biological response to toxicant exposure during pregnancy. Biomarkers, such as levels of specific chemicals in maternal blood or urine, can provide valuable insights into the extent of exposure and its potential impact on maternal and fetal health (Needham et al., 2005)

Research Question:

Against this backdrop, the present study aims to address the following research question: Are there unique profiles based on toxicant exposures and biomarkers among pregnant women? Additionally, how do these latent profiles relate to gestation age at the final prenatal visit? By investigating these questions, we aim to contribute to a better understanding of the complex relationship between toxicant exposure, biomarker response, and gestational outcomes, thereby informing targeted interventions and prenatal care strategies aimed at promoting healthy pregnancies.

Data Overview:

Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

In this research paper, the data analysis was conducted using the statistical software R to explore the relationship between exposure biomarker levels, and gestation age at the final prenatal visit. Here's a summary of our data overview:

The dataset Our dataset consisted of 161 observations, each representing a pregnant individual. We investigated four types of exposures, including phthalates, phenols & parabens, polycyclic aromatic hydrocarbons, and trace metals. These exposures were measured across 38 different variables. The dataset also included risk scores for each type of exposure.

Additionally, we examined six types of endogenous biomarkers, including Cyclooxygenase, Cytochrome P450, Lipoxygenase, Parent Compound, Oxidative Stress, Protein Damage, and Inflammatory biomarkers. These biomarkers were assessed through 61 variables.

To account for potential confounding factors, we included several covariates in our analysis:

1. **Age:** The age of the participants ranged from 23.05 to 47.89 years, with a mean age of 32.58 years. The distribution of ages was as follows: the 1st quartile (Q1) was 29.99 years, the median was 32.84 years, and the 3rd quartile (Q3) was 35.19 years. (Figure 1)
2. **Insurance Status:** Among the participants, 143 individuals had private insurance, while 18 individuals were covered by public health insurance. (Figure 2)
3. **Specific Gravity:** The specific gravity values ranged from 1.001 to 1.031, with a mean of 1.014. The distribution of specific gravity was as follows: the 1st quartile (Q1) was 1.008, the median was 1.014, and the 3rd quartile (Q3) was 1.020. (Figure 3)
4. **Initial Body Mass Index (BMI):** The initial BMI of the participants ranged from 8.494 to 44.086, with a mean BMI of 26.443. The distribution of initial BMI was as follows: the

1st quartile (Q1) was 22.170, the median was 26.185, and the 3rd quartile (Q3) was 30.938. (Figure 4)

5. **Education Level:** The education level of the participants varied, with 23 individuals having a high school degree, 18 individuals attending technical school, 52 individuals attending junior college or some college, and 68 individuals being college graduates. (Figure 5)
6. **Race :** 102 white women, 18 black women, and 41 women from other racial backgrounds. (Figure 6)

These covariates were included in our analysis to control for potential confounding effects and to better understand the relationship between toxicant exposure, biomarker levels, and gestational outcomes.

Our primary response variable of interest was gestation age at the final prenatal visit. This variable served as a measure of pregnancy duration and was used to assess the relationship between toxicant exposure, biomarker levels, and gestational outcomes.

METHODS

We utilized R to apply a range of statistical methods in our analysis. Specifically, we initially conducted Latent Profile Analysis to investigate patterns of toxicant exposures. Subsequently, we experimented with Principal Component Analysis and K Means clustering for the endogenous biomarkers, ultimately opting for Factor Analysis. These methods allowed us to explore the relationships between exposures, biomarkers, covariates, and the response variable. Finally, we employed LASSO regression analysis, offering valuable insights into the influence of these predictors on gestation on final visit. In this section, we will provide an overview of each method employed in our analysis.

Latent Profile Analysis :

It is a statistical procedure used to uncover hidden subgroups within a population by examining patterns of responses to observed variables. It utilizes a probabilistic model to estimate membership probabilities in latent profiles.

Model:

$$\boldsymbol{\sigma_i^2 = \sum_{k=1}^{K} \pi_k (\mu_{ik} - \mu_i)^2 + \sum_{k=1}^{K} \pi_k \, \sigma_{ik}^2}$$

μ_{ik} and σ_{ik} represent profile-specific means and variances for variable ii.

π_k indicates profile density.

Principal component Analysis:

Principal components enable us to condense this set into a smaller number of representative variables that collectively capture the majority of variability present in the original set. PCA (Principal Component Analysis) is characterized as an orthogonal linear transformation applied to a real inner product space. This transformation reconfigures the data into a new coordinate system where the primary principal component represents the highest variance in the data, the second principal component accounts for the subsequent highest variance, and so forth. Essentially, PCA rearranges the data to maximize the spread of variance along the principal axes, facilitating a more efficient representation of the original dataset.

Model:

$$\overbrace{\emptyset_{11},\ldots..,\emptyset_{p1}}^{maximize}\ \{\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\emptyset_{j1}x_{ij})^{2}\ \}$$

$\emptyset_{11},\ldots..,\emptyset_{p1}$ represent the loadings.

n represent the number of observations

p represents the number of predictors.

K-means clustering :

K-means clustering offers a straightforward and efficient method for dividing a dataset into K separate clusters, each distinct and non-overlapping. To execute K-means clustering, we must initially define the desired number of clusters, denoted as K. Subsequently, the K-means algorithm proceeds to allocate each observation in the dataset to precisely one of the K clusters, based on minimizing the distances between observations and cluster centroids. This iterative process continues until convergence, resulting in clearly delineated clusters that best capture the underlying structure of the data.

Model:

$$\overbrace{C_{1},\ldots..C_{K}}^{minimize}\{\sum_{k=1}^{K}\frac{1}{|C_{k}|}\sum_{i,i'\in C_{k}}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^{2}\}$$

$|C_k|$ denotes the number of observations in the kth cluster.

K represents the number of clusters

Factor Analysis:

Factor analysis is used to find latent factors and reduce dimensionality mostly through Principal component analysis and the maximum likelihood method.

Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

PCA is utilized to identify latent factors within the dataset and effectively reduce its dimensionality. By transforming the original variables into a new set of orthogonal variables known as principal components, PCA enables us to capture the most significant sources of variation in the data. This reduction in dimensionality facilitates a more manageable representation of the dataset while preserving as much variance as possible.

The Maximum Likelihood Method is utilized to estimate the parameters of the statistical model governing the data distribution. This method seeks to find the parameter values that maximize the likelihood of observing the given data. In the context of our analysis, the Maximum Likelihood Method is applied in conjunction with PCA to estimate the parameters governing the underlying structure of the data and to identify the principal components that best explain the observed variation.

Model:

$$\widehat{\sum} = \widehat{\Lambda}\widehat{\Lambda}^T + \widehat{\Psi}$$

Variability in data X, is represented by $\sum$.

It's estimate $\widehat{\sum}$ is composed of :

$\widehat{\Lambda}\widehat{\Lambda}^T$: Variability explained by factors (communality).

$\widehat{\Psi}$: Variability not explained by factors (uniqueness)

LASSO:

The LASSO regularization technique serves the dual purpose of minimizing the RSS to enhance predictive accuracy while promoting variable selection by penalizing the absolute size of regression coefficients. This approach enables us to identify and prioritize the most relevant predictors, leading to a more parsimonious and interpretable model.

Model:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ij}\right)^2 + \lambda \sum_{j=i}^{p} |\beta_j|$$

n: Number of observations.

y_i : i^{th} Outcome response.

β_0 : Intercept term.

x_{ij} : Value of the j^{th} predictor for the i^{th} observation.

β_j : Coefficient for the j^{th} predictor.

λ: Regularization parameter, controlling penalty strength.

Data Analysis and Corresponding Results

Latent Profile Analysis on Exposures:

Each of the mentioned toxicant classes - phthalates, phenols and parabens, polycyclic aromatic hydrocarbons (PAHs), and trace metals - essentially consists of chemical compounds sharing similar properties or exposure sources. Following the Latent Profile Analysis with the mclust package, we generated a table presenting three top models effectively clustering the data while disregarding their specific subclasses. The top three models with potential clusters are outlined in (Table 1).

Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Per the results based on the BIC values produced, the model with 4 clusters (VVI,4) has the lowest BIC value, followed by the model with 3 clusters (VVI,3), and then the model with 2 clusters (VVI,2).

Choosing the best number of clusters typically involves selecting the model with the lowest BIC value, as it indicates the best balance between model complexity and fit to the data. In this case, the model with 4 clusters (VVI,4) would be preferred over the others.

To better understand how the clusters were created, a contour plot (Figure 7) was generated based on the clusters, providing a clearer visualization of the four distinct clusters. The contour plot helps visualize how the clusters were formed, offering a clearer understanding of the four distinct clusters. In this context, Dir1, Dir2, and Dir3 represent the primary directions in the original variable space where the data varies the most. The eigenvalues associated with these directions, namely 1.2876, 0.43764, and 0.12843 (Table 2) respectively, indicate how much variance each principal component explains in the data.

The contour plot (Figure 7) displays the data's density projected onto the first two principal components, which capture the most variability in the data. This approach simplifies visualization by condensing the data into two dimensions, making it easier to interpret the clustering patterns and identify clusters visually. Although projecting the data onto fewer dimensions may result in some loss of information, it often provides sufficient insight into the underlying clustering structure. Each unique shape on the plot represents a different latent class.

To explore the relationship between the subfactors of toxicants in the exposure variables, boxplots (Figure 8) were generated for each toxicant cluster. Due to the complexity of visualizing their relationships when plotted together, the boxplots were created individually using a loop function. This approach allowed for a focused study of each toxicant's relationship within its respective cluster. The observed relationships are outlined as follows:

Latent Cluster 1: Primarily characterized by low exposure levels to phthalates, phenols, parabens, and Polycyclic Aromatic Hydrocarbons (PAHs).

Latent Cluster 2: Typically exhibits average exposure levels to phthalates, phenols, parabens, Polycyclic Aromatic Hydrocarbons (PAHs), and trace metals.

Latent Cluster 3: Predominantly associated with high exposure levels to phthalates, phenols, parabens, Polycyclic Aromatic Hydrocarbons (PAHs), and trace metals.

Latent Cluster 4: Characterized by very low exposure levels to trace metals.

Factor Analysis on Biomarkers:

The endogenous biomarkers Factor comprises six subfactors associated with biological pathways, including the Cyclooxygenase Pathway, Cytochrome P450 Pathway, Lipoxygenase Pathway, Parent Compound, Oxidative Stress, Protein Damage, and Inflammatory responses. These highlighted pathways and biomarkers are integral to human physiology and health, playing crucial roles in various physiological functions and disease processes. Understanding the significance of pathways such as Cyclooxygenase, Cytochrome P450, Lipoxygenase, and markers like Oxidative Stress and Inflammatory responses is essential for diagnosing conditions and developing targeted treatments aimed at maintaining overall health and well-being. The data underwent factor analysis using the psych package to identify latent factors. We utilized a scree (Figure 9) plot to identify the number of factors with the greatest variability. Employing the elbow method, we experimented with different numbers of factors at the bend of the scree plot.

Ultimately, we discovered that 18 factors, following the Promax method for factor correlation, elucidated roughly 80% of the variance. The scree plot (Figure 9) serves to pinpoint the optimal number of factors, maximizing coverage. It's a visual aid where the actual elbow signifies the minimum necessary factors (AF). Meanwhile, the maximum feasible factors (OC) are determined by the last eigenvalue being ≥ 1. Values greater than 1 are preferred, while those below 1, especially, are less desirable. The green triangles marks separate the eigenvalues less than 1 and those greater one and the redline the shows which eigenvalues are lesser than one, so below it is the values less than 1 The green triangles signify the separation between eigenvalues less than 1 and those greater than one, while the red line indicates eigenvalues below 1. So, values below the red line are those less than 1.

Finally, Lasso regression was employed on the 18 factors derived from the factor analysis on biomarkers, the 4 latent clusters concerning toxicants, the 6 covariates, and the risk scores. This aimed to select the most significant predictors for conducting a multiple regression analysis. Additionally, the specific gravity (sg) variable was omitted since all observations fell within the normal range of 1.005 to 1.030, rendering it unnecessary. Furthermore, the intercept was excluded from the model, as the dataset encompassed all potential exposures and biomarkers.

Results

The conclusion drawn is that initial BMI, Factor 10, and education are statistically significant at the 5% significance level in relation to gestational age at the final visit. Furthermore, the high R-squared value of 0.957 suggests that 95.7% of the variability in gestational age at the final visit is explained by the predictors incorporated into the model.

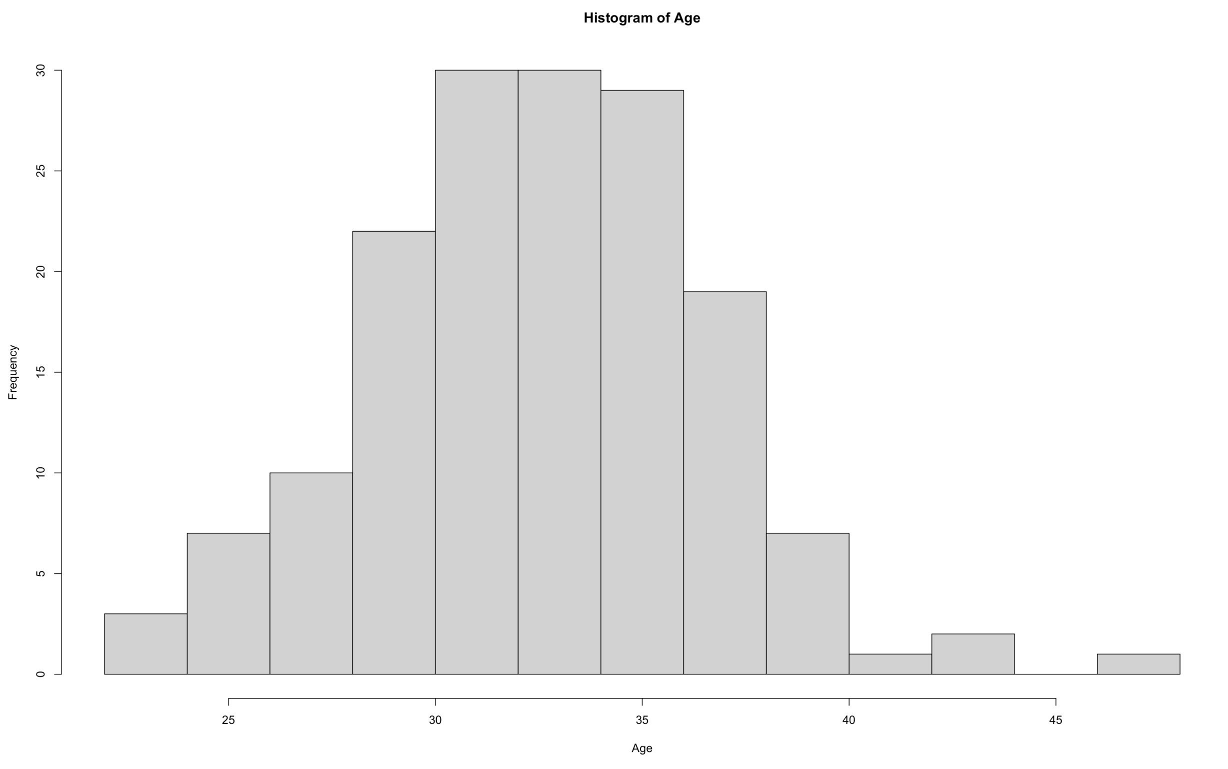
Limitations

K-means clustering, and Principal Component Analysis (PCA) were attempted on the Endogenous Biomarkers. However, a limited number of factors failed to adequately explain variance. Nearly half the number of predictors is required to sufficiently describe a decent amount variance in the data.

Another challenge encountered was the inability to determine the relationships between clusters due to the complexity of the data.

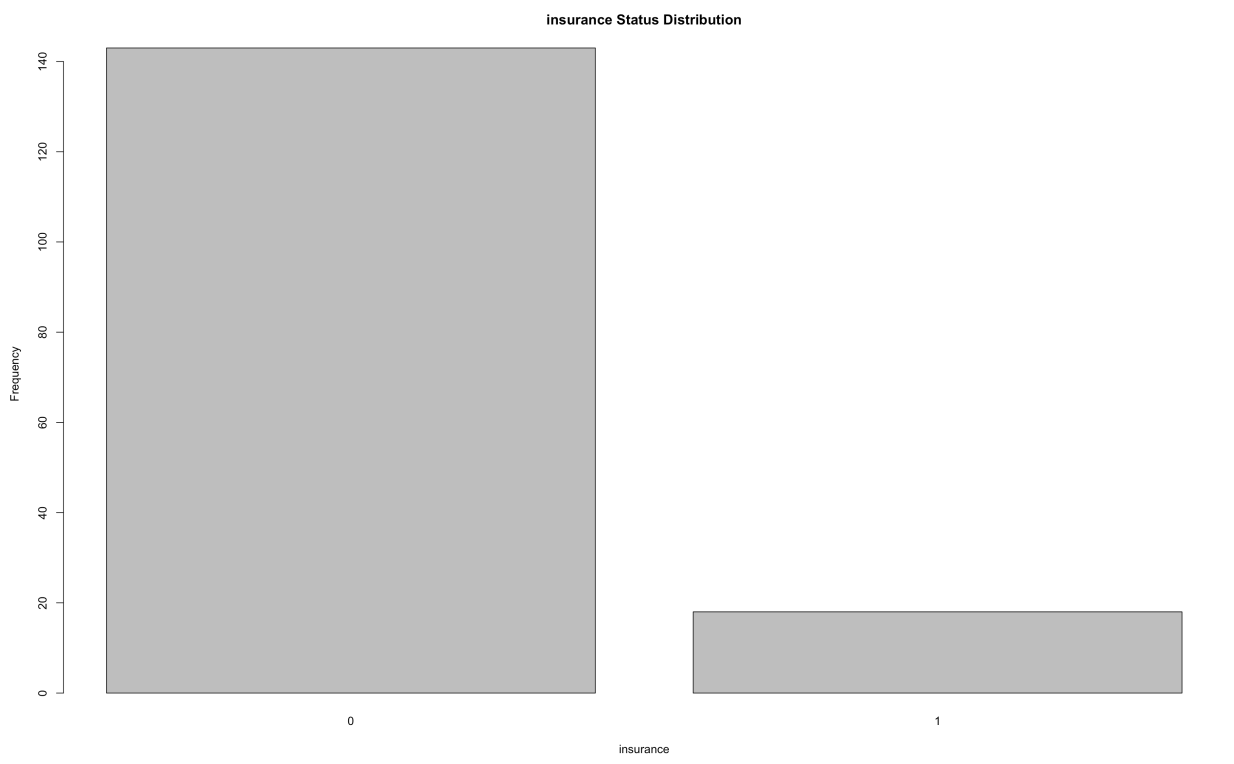
Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Figure 1



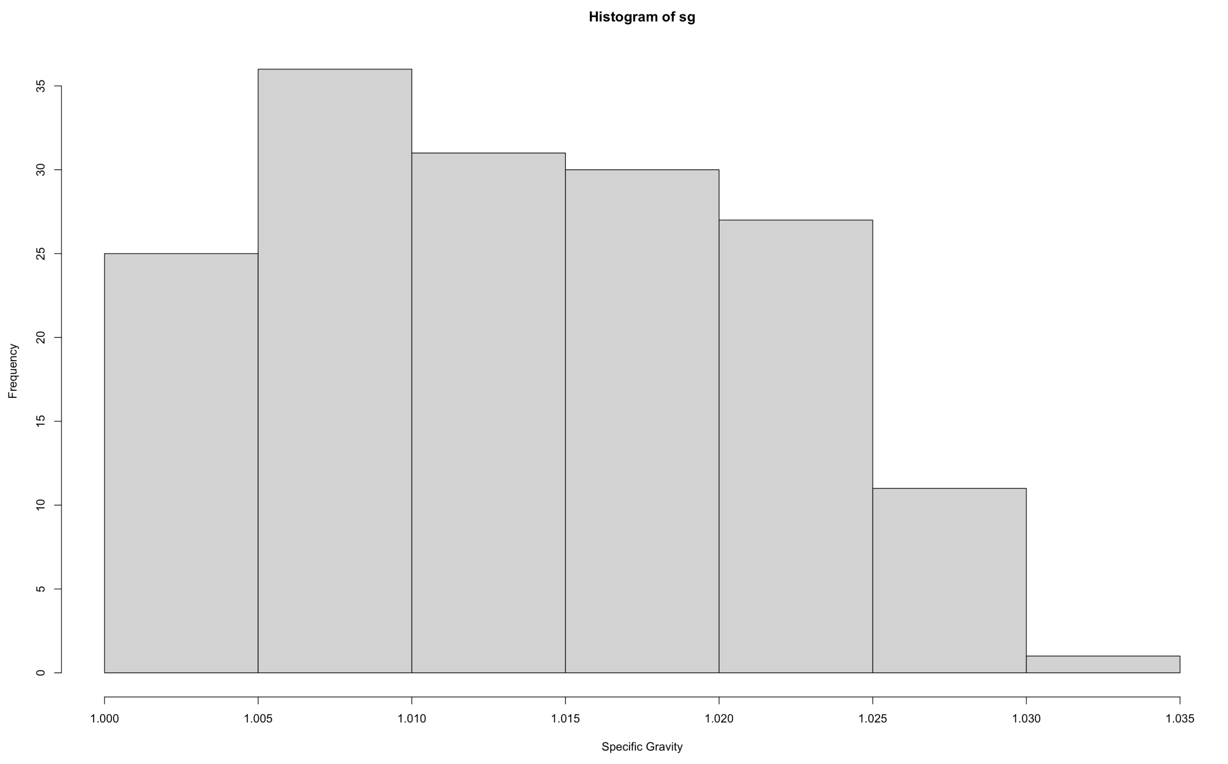
Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Figure 2



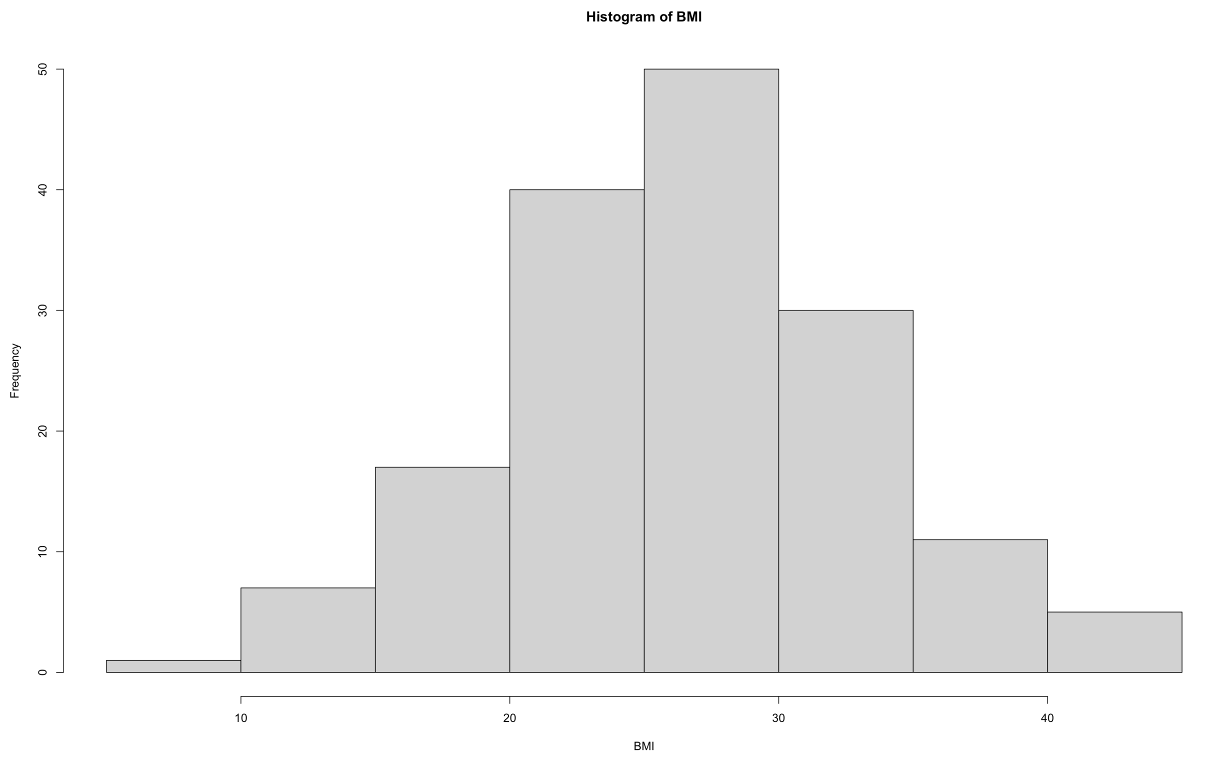
Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Figure 3



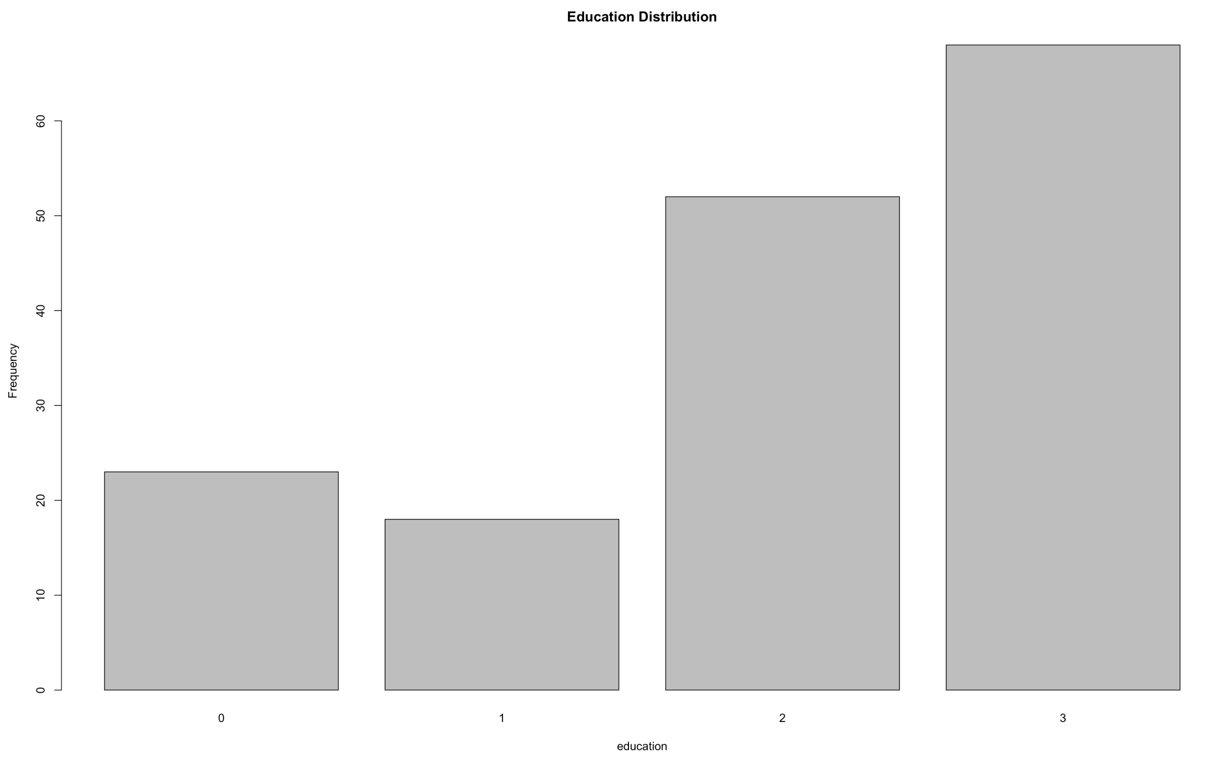
Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Figure 4



Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Figure 5



Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Figure 6

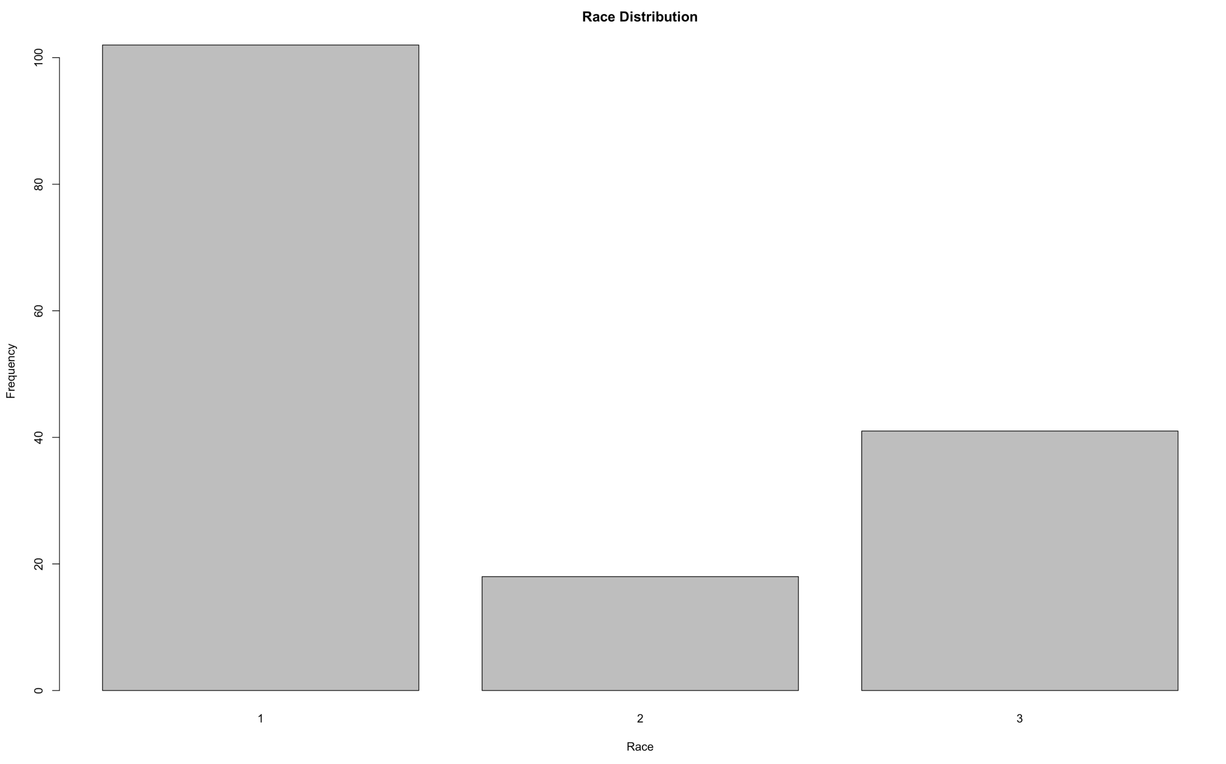


Figure 7

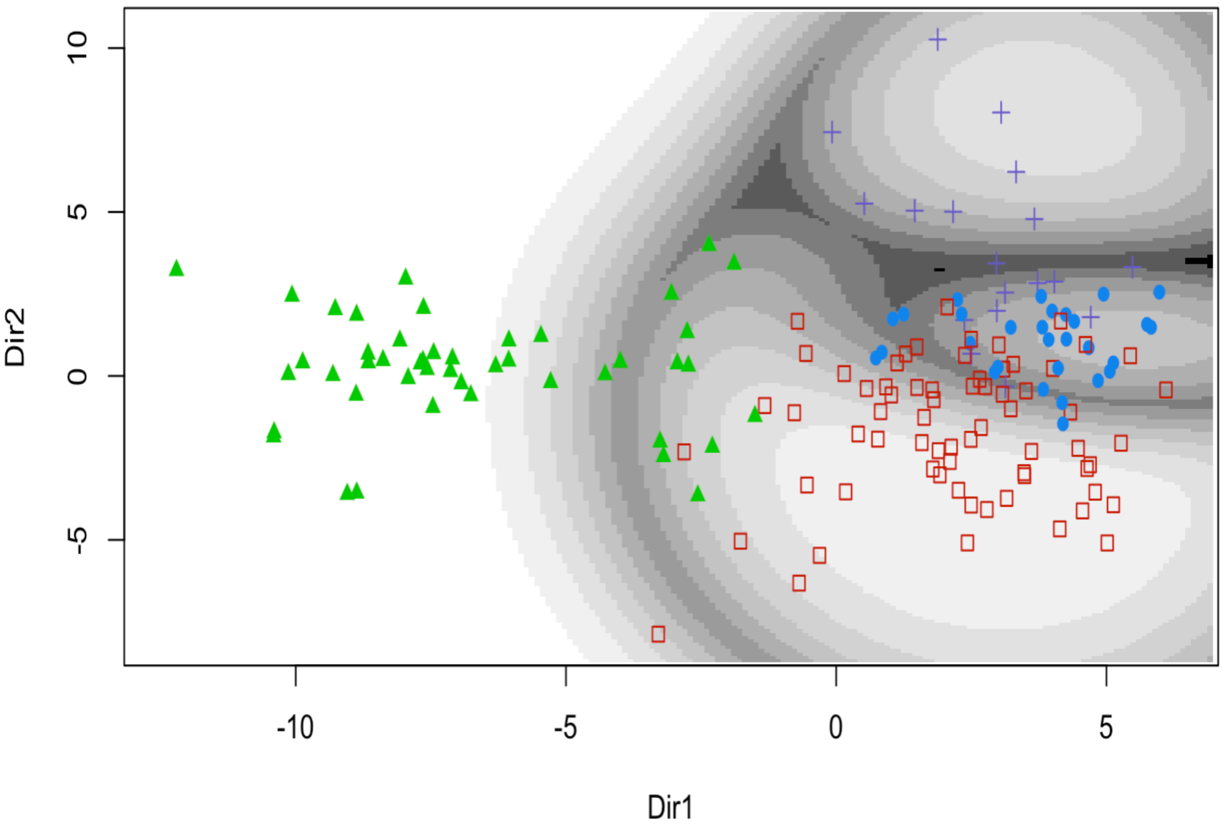
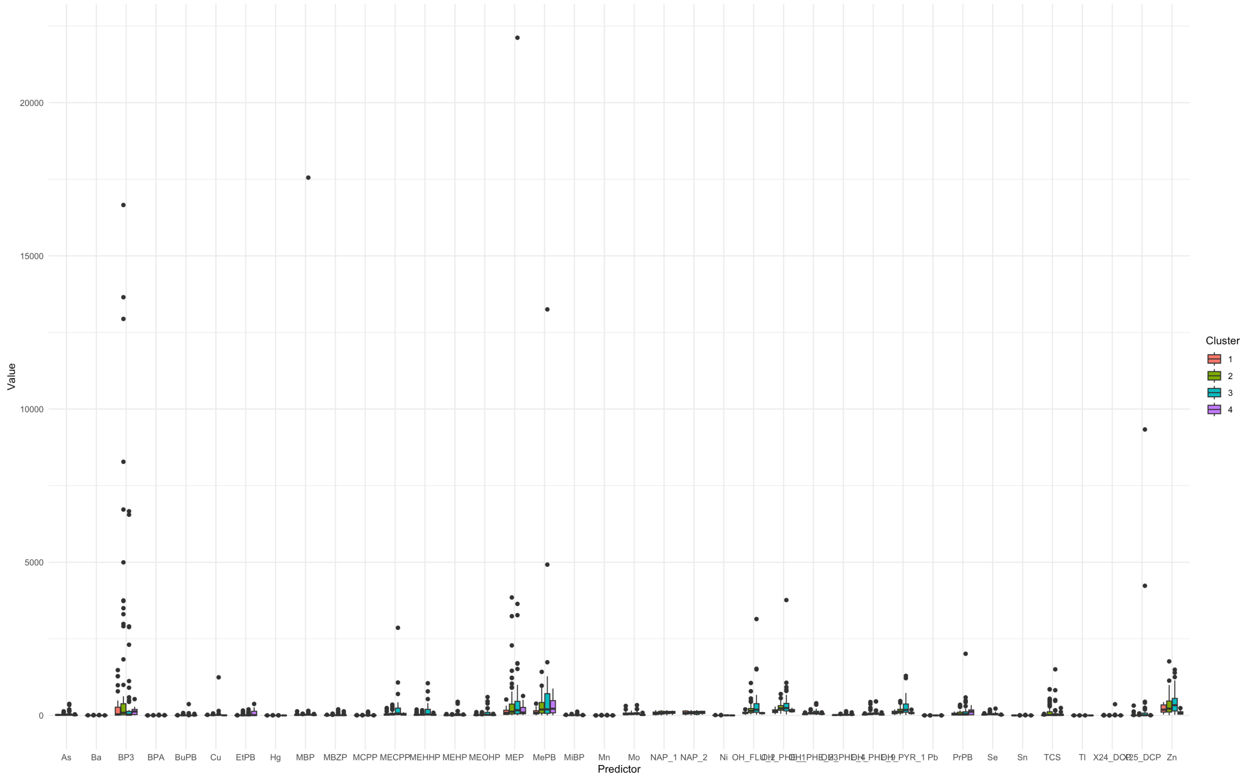
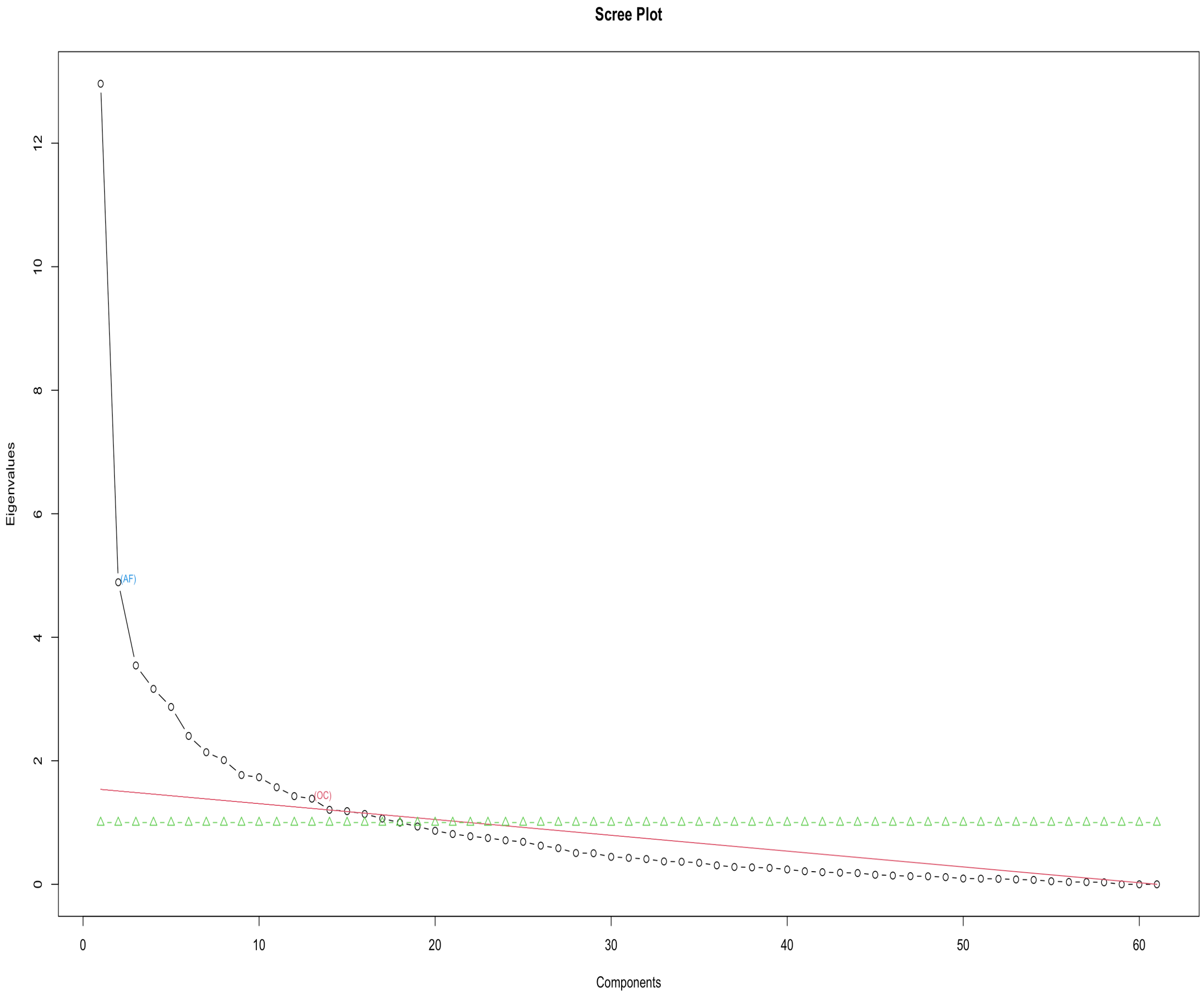


Figure 8


20000
15000
10000
5000
0
Value
As
Ba
BP3
BPA
BuPB
Cu
EtPB
Hg
MBP
MBZP
MCPP
MECPP
MEHHP
MEHP
MEOHP
MEP
MePB
MiBP
Mn
Mo
NAP_1
NAP_2
Ni
OH_FLU
PHE
PHE
PHE
PHE
PYR_1
Pb
PrPB
Se
Sn
TCS
Tl
X24_DCP
X25_DCP
Zn
Predictor
Cluster
1
2
3
4

Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Figure 9



Exploring Latent Profiles of Toxicant Exposure and Biomarker Levels Among Pregnant Women and Their Relationship with Gestation Age on Final Visit.

Table 1	VVI,4	VVI,3	VVI,2
BIC	-56433.19	-57264.568	-58096.586
BIC diff	0.00	-831.376	-1663.395

Table 2	Dir1	Dir2	Dir3
Eigenvalues	1.2876	0.43764	0.12843
Cum. %	69.4622	93.07157	100.00000

Table3 Predictor	Coefficient	Std. Error	T-value	P-value
t1bmi	1.09165	0.04642	23.515	< 2e-16
education	3.78649	0.55777	6.789	2.51e-10
Factor10	0.86147	0.35384	2.435	0.0161

REFERENCES

Braun, J. M., Lanphear, B. P., & Vigoren, E. M. (2016). Early-life exposure to EDCs: role in childhood obesity and neurodevelopment. Nature Reviews Endocrinology, 13(3), 161-173.

Grandjean, P., & Landrigan, P. J. (2006). Developmental neurotoxicity of industrial chemicals. The Lancet, 368(9553), 2167-2178.

Hartmann, K., Krois, J., & Rudolph, A. (2023). *Statistics and Geodata Analysis using R (SOGA-R).* Department of Earth Sciences, Freie Universitaet Berlin.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning.* Accessed April 28, 2024, from https://www.statlearning.com/.

Needham, L. L., Barr, D. B., Caudill, S. P., Pirkle, J. L., Turner, W. E., Osterloh, J., ... & Sampson, E. J. (2005). Concentrations of environmental chemicals associated with neurodevelopmental effects in US population. Neurotoxicology, 26(4), 531-545.

Spurk, D., Hirschi, A., Wang, M., Valero, D. C., & Kauffeld, S. (2020). *Latent profile analysis: A review and 'how to' guide of its application within vocational behavior research. Journal of Vocational Behavior, 120*, 103445.

Wigle, D. T., Arbuckle, T. E., & Walker, M. (2008). Epidemiologic evidence of relationships between reproductive and child health outcomes and environmental chemical contaminants. Journal of Toxicology and Environmental Health, Part B, 11(5-6), 373-517.