Supplemental Digital Content II

Title: Per- and Polyfluoroalkyl Substances in Drinking Water and Birthweight in the US: A County-level Study

Authors: Yachen Zhu[1], Scott M. Bartell[1,2]

Author affiliations:
1. Program in Public Health, University of California, Irvine, CA 92697-3957, USA
2. Department of Statistics, University of California, Irvine, CA 92697-1250, USA

Corresponding Author:
Yachen Zhu
Program in Public Health
Anteater Instruction and Research Building
Irvine, CA 92697-3957
USA
Email: yachenz1@uci.edu

Using multiple-stratified average birthweights data in weighted regression models (group-level analysis) produces equivalent slope and theoretical variance (or standard error) of effect estimate to those that would be obtained from using individual-level data in unweighted regression models (individual-level analysis). However, the variance (or standard error) of effect estimate reported by statistical software (such as R) would be different for these two types of analyses. In addition to the proof of equivalent regression slope for the two types of analyses, we present the corrected variance formulas here.

Suppose there are n individuals, each with birthweight $Y_i,\ i = 1, ..., n$, and p ($p \geq 1$) variables in the regression model. For individual $i, X_{i1}, X_{i2}, ...,$ and X_{ip} are the values of the p explanatory variables. Suppose the n individuals are divided into m groups ($m < n$). For group j ($j = 1, 2, ..., m$), there are n_j members, $\sum_{j=1}^{m} n_j = n$. In each group, the individuals share the same vector of explanatory variables $\boldsymbol{X_i^T} = (1, X_{i1}, \cdots, X_{ip})$. To simplify the notation, let $\boldsymbol{Z_j^T} = (1, Z_{j1}, \cdots, Z_{jp})$ be the vector for group j's explanatory variables. Let $\boldsymbol{U^T} = (U_1, U_2, \cdots, U_m)$ denote the mean of the response variable for the m groups. $U_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$ is the average birthweight for group j.

$$\mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n_1,1} & \cdots & X_{n_1,p} \\ 1 & X_{n_1+1,1} & \cdots & X_{n_1+1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n_1+n_2,1} & \cdots & X_{n_1+n_2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1+\sum_{j=1}^{m-1} n_j,1} & \cdots & X_{1+\sum_{j=1}^{m-1} n_j,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix}_{n\times(1+p)} \triangleq \begin{pmatrix} 1 & Z_{1,1} & \cdots & Z_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{1,1} & \cdots & Z_{1,p} \\ 1 & Z_{2,1} & \cdots & Z_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{2,1} & \cdots & Z_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{m,1} & \cdots & Z_{m,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{m,1} & \cdots & Z_{m,p} \end{pmatrix}_{n\times(1+p)}$$

Using individual-level data, we have

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{X_1^T} \\ \boldsymbol{X_2^T} \\ \vdots \\ \boldsymbol{X_n^T} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$= \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_{11} + \cdots + \beta_p X_{1p} + \varepsilon_1 \\ \beta_0 + \beta_1 X_{21} + \cdots + \beta_p X_{2p} + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \cdots + \beta_p X_{np} + \varepsilon_n \end{pmatrix}$$

Using grouped data, we have

$$\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_m \end{pmatrix} = \mathbf{Z} \cdot \tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}} = \begin{pmatrix} \boldsymbol{Z_1^T} \\ \boldsymbol{Z_2^T} \\ \vdots \\ \boldsymbol{Z_m^T} \end{pmatrix} \cdot \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_p \end{pmatrix} + \begin{pmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \vdots \\ \tilde{\varepsilon}_m \end{pmatrix}$$

$$= \begin{pmatrix} 1 & Z_{11} & \cdots & Z_{1p} \\ 1 & Z_{21} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{m1} & \cdots & Z_{mp} \end{pmatrix} \cdot \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_p \end{pmatrix} + \begin{pmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \vdots \\ \tilde{\varepsilon}_m \end{pmatrix} = \begin{pmatrix} \tilde{\beta}_0 + \tilde{\beta}_1 Z_{11} + \cdots + \tilde{\beta}_p Z_{1p} + \tilde{\varepsilon}_1 \\ \tilde{\beta}_0 + \tilde{\beta}_1 Z_{21} + \cdots + \tilde{\beta}_p Z_{2p} + \tilde{\varepsilon}_2 \\ \vdots \\ \tilde{\beta}_0 + \tilde{\beta}_1 Z_{m1} + \cdots + \tilde{\beta}_p Z_{mp} + \tilde{\varepsilon}_m \end{pmatrix}$$

$\mathbf{Z} = \begin{pmatrix} 1 & Z_{11} & \cdots & Z_{1p} \\ 1 & Z_{21} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n1} & \cdots & Z_{mp} \end{pmatrix}$ only includes the unique vectors in $\mathbf{X}$.

(1) Assuming constant variance for individual birthweight, i.e., $Var(Y_i) = \sigma^2, i = 1, \cdots, n$
Using individual-level data and according to the ordinary least squares

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_1 =& (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
=& \left[\begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix} \cdot \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \\
=& \begin{pmatrix} n & \sum_{i=1}^n X_{i1} & \cdots & \sum_{i=1}^n X_{ip} \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \cdots & \sum_{i=1}^n X_{i1}X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ip} & \sum_{i=1}^n X_{ip}X_{i1} & \cdots & \sum_{i=1}^n X_{ip}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{i1}Y_i \\ \vdots \\ \sum_{i=1}^n X_{ip}Y_i \end{pmatrix} \\
=& \begin{pmatrix} n & \sum_{j=1}^m n_j Z_{j1} & \cdots & \sum_{j=1}^m n_j Z_{jp} \\ \sum_{j=1}^m n_j Z_{j1} & \sum_{j=1}^m n_j Z_{j1}^2 & \cdots & \sum_{j=1}^m n_j Z_{j1}Z_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^m n_j Z_{jp} & \sum_{j=1}^m n_j Z_{jp}Z_{j1} & \cdots & \sum_{j=1}^m n_j Z_{jp}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^m n_j U_j \\ \sum_{j=1}^m n_j Z_{j1}U_j \\ \vdots \\ \sum_{j=1}^m n_j Z_{jp}U_j \end{pmatrix}
\end{aligned}
$$

Using grouped data and according to the weighted least squares using the number of births in each stratum for the weights

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_2 =& (\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}_1\mathbf{U} \\
=& \left[\begin{pmatrix} 1 & 1 & \cdots & 1 \\ Z_{11} & Z_{21} & \cdots & Z_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1p} & Z_{2p} & \cdots & Z_{mp} \end{pmatrix} \cdot \begin{pmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_m \end{pmatrix} \cdot \begin{pmatrix} 1 & Z_{11} & \cdots & Z_{1p} \\ 1 & Z_{21} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n1} & \cdots & Z_{mp} \end{pmatrix} \right]^{-1} \\
& \cdot \begin{pmatrix} 1 & 1 & \cdots & 1 \\ Z_{11} & Z_{21} & \cdots & Z_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1p} & Z_{2p} & \cdots & Z_{mp} \end{pmatrix} \begin{pmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_m \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_m \end{pmatrix} \\
=& \begin{pmatrix} n & \sum_{j=1}^m n_j Z_{j1} & \cdots & \sum_{j=1}^m n_j Z_{jp} \\ \sum_{j=1}^m n_j Z_{j1} & \sum_{j=1}^m n_j Z_{j1}^2 & \cdots & \sum_{j=1}^m n_j Z_{j1}Z_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^m n_j Z_{jp} & \sum_{j=1}^m n_j Z_{jp}Z_{j1} & \cdots & \sum_{j=1}^m n_j Z_{jp}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^m n_j U_j \\ \sum_{j=1}^m n_j Z_{j1}U_j \\ \vdots \\ \sum_{j=1}^m n_j Z_{jp}U_j \end{pmatrix}
\end{aligned}
$$

where $\mathbf{W}_1 = \begin{pmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_m \end{pmatrix}_{m \times m}$ is the weight matrix.

Therefore, we have

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_2$$

Therefore, using the number of births in each stratum for the weights produces equivalent regression estimates to those that would be obtained from using unweighted multiple linear regression with the individual-level data.

In addition, because we assume that the individual birthweight has a constant variance σ^2, then $Var(U_j) = Var(\frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}) = \frac{1}{n_j^2}\sum_{i=1}^{n_j} Var(Y_{ij}) = \frac{n_j}{n_j^2}\sigma^2 = \frac{\sigma^2}{n_j}$. Because we use the number of births in each stratum (n_j) as the weight in the regression using grouped data, and different observations are independent of each other (Independence Assumption of Regression), then $Var(\mathbf{U}) = diag(\frac{\sigma^2}{n_j}) = \sigma^2 diag(\frac{1}{n_j}) = \sigma^2\mathbf{W}_1^{-1}$. Therefore, we have

$$\begin{aligned}
Var(\hat{\boldsymbol{\beta}}_1) =& Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}) \\
=& (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot Var(\mathbf{Y}) \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
=& (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \sigma^2\mathbf{I} \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
=& \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
=& \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

$$\begin{aligned}
Var(\hat{\boldsymbol{\beta}}_2) =& Var((\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}_1\mathbf{U}) \\
=& (\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}_1 \cdot Var(\mathbf{U}) \cdot \mathbf{W}_1^T\mathbf{Z}(\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1} \\
=& (\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}_1 \cdot \sigma^2\mathbf{W}_1^{-1} \cdot \mathbf{W}_1^T\mathbf{Z}(\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1} \\
=& \sigma^2(\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}_1^T\mathbf{Z}(\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1} \\
=& \sigma^2(\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1} \ \text{(because } \mathbf{W}_1^T = \mathbf{W}_1\text{)}
\end{aligned}$$

From the derivation on page 3, we know that $(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{Z}^T\mathbf{W}_1\mathbf{Z})^{-1}$, therefore in theory we have $Var(\hat{\boldsymbol{\beta}}_1) = Var(\hat{\boldsymbol{\beta}}_2)$. However, in statistical packages such as R, σ^2 in $Var(\hat{\boldsymbol{\beta}}_1)$ is estimated by the mean squared error $MSE = \frac{SSE}{n-p-1}$; while in $Var(\hat{\boldsymbol{\beta}}_2)$, σ^2 is estimated by the weighted mean squared error $WMSE = \frac{WSSE}{m-p-1}$, $WSSE$ is the weighted sum of squared errors. Therefore,

$$\begin{aligned}
& \frac{Var(\hat{\boldsymbol{\beta}}_1)}{Var(\hat{\boldsymbol{\beta}}_2)} = \frac{SSE/(n-p-1)}{WSSE/(m-p-1)} = \frac{SSE}{WSSE} \cdot \frac{m-p-1}{n-p-1} \\
\Rightarrow\ & Var(\hat{\boldsymbol{\beta}}_1) = \frac{SSE}{WSSE} \cdot \frac{m-p-1}{n-p-1} \cdot Var(\hat{\boldsymbol{\beta}}_2) \\
\Rightarrow\ & \hat{se}(\hat{\boldsymbol{\beta}}_1) = \sqrt{\frac{SSE}{WSSE} \cdot \frac{m-p-1}{n-p-1}} \cdot \hat{se}(\hat{\boldsymbol{\beta}}_2)
\end{aligned}$$

Now we know that weighted regression using multiple-stratified birthweight data produces the same regression slope as individual-level analysis using the same variables ($\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_2$), and in our analysis the individuals in each group share the same vector of explanatory variables, therefore the estimated birthweight for individual i in group j is equal to the estimated average birthweight for group j:

$$\hat{Y}_{ij} = \hat{U}_j$$

The sum of squared errors (SSE) from unweighted regression using individual-level data is

$$\begin{aligned}
SSE &= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\hat{Y}_{ij})^2 \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-U_j+U_j-\hat{Y}_{ij})^2 \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-U_j)^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(U_j-\hat{Y}_{ij})^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}2(Y_{ij}-U_j)(U_j-\hat{Y}_{ij}) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-U_j)^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(U_j-\hat{U}_j)^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}2(Y_{ij}-U_j)(U_j-\hat{U}_j) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-U_j)^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(U_j-\hat{U}_j)^2+\sum_{j=1}^{m}2(U_j-\hat{U}_j)\sum_{i=1}^{n_j}(Y_{ij}-U_j) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-U_j)^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(U_j-\hat{U}_j)^2+\sum_{j=1}^{m}2(U_j-\hat{U}_j)\left(\sum_{i=1}^{n_j}Y_{ij}-n_jU_j\right) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-U_j)^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(U_j-\hat{U}_j)^2+\sum_{j=1}^{m}2(U_j-\hat{U}_j)\cdot 0 \\
&= \sum_{j=1}^{m}n_j\sigma_j^2+\sum_{j=1}^{m}n_j(U_j-\hat{U}_j)^2
\end{aligned}$$

because the population variance of birthweight in group j is $\sigma_j^2=\frac{1}{n_j}\sum_{i=1}^{n_j}(Y_{ij}-U_j)^2$.
In the above formula of SSE, the second term is the weighted sum of squared errors from group-level regression weighted by group size n_j, i.e.,

$$WSSE=\sum_{j=1}^{m}n_j\tilde{\varepsilon}_j^2=\sum_{j=1}^{m}n_j(U_j-\hat{U}_j)^2$$

where $\tilde{\varepsilon}_j=U_j-\hat{U}_j$ is the residual of group-level analysis, which can be extracted from the fitted regression models. Therefore,

$$SSE=\sum_{j=1}^{m}n_j\sigma_j^2+WSSE$$

$$\begin{aligned}
\hat{se}(\hat{\boldsymbol{\beta}}_1) &= \sqrt{\frac{SSE}{WSSE}\cdot\frac{m-p-1}{n-p-1}}\cdot\hat{se}(\hat{\boldsymbol{\beta}}_2) \\
&= \sqrt{\frac{\sum_{j=1}^{m}n_j\sigma_j^2+WSSE}{WSSE}\cdot\frac{m-p-1}{n-p-1}}\cdot\hat{se}(\hat{\boldsymbol{\beta}}_2)
\end{aligned}$$

For the models that use the percentage of water measurements with PFAS detection as the exposure metric, the standard errors in the crude models were reduced by 92%, and the standard errors in the adjusted and adjusted co-exposure models were reduced by 30% in unweighted individual-level analysis compared with the group-level analysis weighted by group size.

For the models that use the population-weighted average PFAS concentration as the exposure metric, the standard errors in the crude models were reduced by 92%, and the standard errors in the adjusted and adjusted co-exposure models were reduced by 34-35% in unweighted individual-level analysis compared with the group-level analysis weighted by group size.

```
### R function to correct the confidence interval for regression parameter
### in unweighted individual-level regression

f1 <- function(data = data, mod = mod){
# input data and group-level regression model weighted by group size
require(broom)
n <- sum(data$Births)
# sum up the number of births in each group to get the total number of births
m <- nrow(data)
# number of groups
wsse.group <- sum(data$Births*(mod$residuals^2))
# weighted SSE for group-level regression weighted by group size
p <- nrow(tidy(mod)) - 1
# number of regression parameters except for the intercept
frac <- sqrt( (sum(data$Births*data$SD^2) + wsse.group)*(m-p-1)/((n-p-1)*wsse.group) )
# the coefficient of the standard error of regression parameter for
# group-level analysis weighted by group size to get the standard error of
# regression parameter for individual-level unweighted analysis
lo <- tidy(mod)$estimate[2] - frac * tidy(mod)$std.error[2] * qnorm(0.975)
# lower limit of confidence interval of regression parameter for individual-level analysis
hi <- tidy(mod)$estimate[2] + frac * tidy(mod)$std.error[2] * qnorm(0.975)
# upper limit of confidence interval of regression parameter for
# individual-level unweighted analysis
print(c(lo, hi))
# confidence interval of regression coefficient for individual-level unweighted analysis
print(1-frac)
# percentage of reduction in standard error compared with
# the group-level analysis weighted by group size
}
```

(2) Assuming non-constant variance for individual birthweight, with each group having a different variance for the birthweight, i.e., $Var(Y_{ij}) = \sigma_j^2,\ j = 1, \cdots, m$.
Using individual-level data in inverse-variance weighted regression (here we use the standard deviation of birthweight in the group to represent the standard deviation of birthweight for the individual). The weight matrix is:

$$\mathbf{W}_2 = Var(\mathbf{Y})^{-1} = \begin{pmatrix} \frac{1}{Var(Y_1)} & & & \\ & \frac{1}{Var(Y_2)} & & \\ & & \ddots & \\ & & & \frac{1}{Var(Y_n)} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1^2} & & & & & & \\ & \ddots & & & & & \\ & & \frac{1}{\sigma_1^2} & & & & \\ & & & \ddots & & & \\ & & & & \frac{1}{\sigma_m^2} & & \\ & & & & & \ddots & \\ & & & & & & \frac{1}{\sigma_m^2} \end{pmatrix}_{n \times n}$$

According to the weighted least squares (WLS), we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_3 =& (\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_2\mathbf{Y} \\
=& \left[\begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \\ Z_{11} & \cdots & Z_{11} & \cdots & Z_{m1} & \cdots & Z_{m1} \\ Z_{12} & \cdots & Z_{12} & \cdots & Z_{m2} & \cdots & Z_{m2} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Z_{1p} & \cdots & Z_{1p} & \cdots & Z_{mp} & \cdots & Z_{mp} \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1^2} & & & & & & \\ & \ddots & & & & & \\ & & \frac{1}{\sigma_1^2} & & & & \\ & & & \ddots & & & \\ & & & & \frac{1}{\sigma_m^2} & & \\ & & & & & \ddots & \\ & & & & & & \frac{1}{\sigma_m^2} \end{pmatrix} \cdot \begin{pmatrix} 1 & Z_{1,1} & \cdots & Z_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{1,1} & \cdots & Z_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{m,1} & \cdots & Z_{m,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{m,1} & \cdots & Z_{m,p} \end{pmatrix} \right]^{-1} \\
& \cdot \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \\ Z_{11} & \cdots & Z_{11} & \cdots & Z_{m1} & \cdots & Z_{m1} \\ Z_{12} & \cdots & Z_{12} & \cdots & Z_{m2} & \cdots & Z_{m2} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Z_{1p} & \cdots & Z_{1p} & \cdots & Z_{mp} & \cdots & Z_{mp} \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1^2} & & & & & & \\ & \ddots & & & & & \\ & & \frac{1}{\sigma_1^2} & & & & \\ & & & \ddots & & & \\ & & & & \frac{1}{\sigma_m^2} & & \\ & & & & & \ddots & \\ & & & & & & \frac{1}{\sigma_m^2} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \\
=& \begin{pmatrix} \sum_{j=1}^m \frac{n_j}{\sigma_j^2} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1} & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp} \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1}^2 & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1}Z_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp}Z_{j1} & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp}^2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sigma_1^2} & \cdots & \frac{1}{\sigma_1^2} & \cdots & \frac{1}{\sigma_m^2} & \cdots & \frac{1}{\sigma_m^2} \\ \frac{Z_{11}}{\sigma_1^2} & \cdots & \frac{Z_{11}}{\sigma_2^2} & \cdots & \frac{Z_{m1}}{\sigma_m^2} & \cdots & \frac{Z_{m1}}{\sigma_m^2} \\ \frac{Z_{12}}{\sigma_1^2} & \cdots & \frac{Z_{12}}{\sigma_2^2} & \cdots & \frac{Z_{m2}}{\sigma_m^2} & \cdots & \frac{Z_{m2}}{\sigma_m^2} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{Z_{1p}}{\sigma_1^2} & \cdots & \frac{Z_{1p}}{\sigma_2^2} & \cdots & \frac{Z_{mp}}{\sigma_m^2} & \cdots & \frac{Z_{mp}}{\sigma_m^2} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \\
=& \begin{pmatrix} \sum_{j=1}^m \frac{n_j}{\sigma_j^2} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1} & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp} \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1}^2 & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1}Z_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp}Z_{j1} & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^m \frac{n_j}{\sigma_j^2} U_j \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1}U_j \\ \vdots \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp}U_j \end{pmatrix}
\end{aligned}$$

Using grouped data in inverse-variance weighted regression, the weight matrix is

$$\tilde{\mathbf{W}}_2 = Var(\mathbf{U})^{-1} = \begin{pmatrix} \frac{1}{Var(U_1)} & & & \\ & \frac{1}{Var(U_2)} & & \\ & & \ddots & \\ & & & \frac{1}{Var(U_m)} \end{pmatrix} = \begin{pmatrix} \frac{n_1}{\sigma_1^2} & & & \\ & \frac{n_2}{\sigma_2^2} & & \\ & & \ddots & \\ & & & \frac{n_m}{\sigma_m^2} \end{pmatrix}_{m\times m}$$

because $Var(U_j) = Var(\frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}) = \frac{1}{n_j^2}\sum_{i=1}^{n_j} Var(Y_{ij}) = \frac{\sigma_j^2}{n_j},\ j = 1, \cdots, m.$
According to the weighted least squares (WLS), we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_4 =& (\mathbf{Z}^T\tilde{\mathbf{W}}_2\mathbf{Z})^{-1}\mathbf{Z}^T\tilde{\mathbf{W}}_2\mathbf{U} \\
=& \left[\begin{pmatrix} 1 & 1 & \cdots & 1 \\ Z_{11} & Z_{21} & \cdots & Z_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1p} & Z_{2p} & \cdots & Z_{mp} \end{pmatrix} \cdot \begin{pmatrix} \frac{n_1}{\sigma_1^2} & & & \\ & \frac{n_2}{\sigma_2^2} & & \\ & & \ddots & \\ & & & \frac{n_m}{\sigma_m^2} \end{pmatrix} \cdot \begin{pmatrix} 1 & Z_{11} & \cdots & Z_{1p} \\ 1 & Z_{21} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n1} & \cdots & Z_{mp} \end{pmatrix} \right]^{-1} \\
& \cdot \begin{pmatrix} 1 & 1 & \cdots & 1 \\ Z_{11} & Z_{21} & \cdots & Z_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1p} & Z_{2p} & \cdots & Z_{mp} \end{pmatrix} \begin{pmatrix} \frac{n_1}{\sigma_1^2} & & & \\ & \frac{n_2}{\sigma_2^2} & & \\ & & \ddots & \\ & & & \frac{n_m}{\sigma_m^2} \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_m \end{pmatrix} \\
=& \begin{pmatrix} \sum_{j=1}^m \frac{n_j}{\sigma_j^2} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1} & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp} \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1}^2 & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1}Z_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp} & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp}Z_{j1} & \cdots & \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^m \frac{n_j}{\sigma_j^2} U_j \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{j1}U_j \\ \vdots \\ \sum_{j=1}^m \frac{n_j}{\sigma_j^2} Z_{jp}U_j \end{pmatrix}
\end{aligned}$$

We have

$$\hat{\boldsymbol{\beta}}_4 = \hat{\boldsymbol{\beta}}_3$$

Therefore, using grouped data in inverse-variance weighted regression produces equivalent estimates to those that would be obtained from individual-level weighted regression allowing for heteroscedasticity.

Because $\mathbf{W}_2 = Var(\mathbf{Y})^{-1}$ and $\tilde{\mathbf{W}}_2 = Var(\mathbf{U})^{-1}$,

$$\begin{aligned}
Var(\hat{\boldsymbol{\beta}}_3) &= Var((\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_2\mathbf{Y}) \\
&= (\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_2 \cdot Var(\mathbf{Y}) \cdot \mathbf{W}_2\mathbf{X}(\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_2 \cdot \mathbf{W}_2^{-1} \cdot \mathbf{W}_2\mathbf{X}(\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_2\mathbf{X}(\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1} \\
\text{Similarly, } Var(\hat{\boldsymbol{\beta}}_4) &= (\mathbf{Z}^T\tilde{\mathbf{W}}_2\mathbf{Z})^{-1}
\end{aligned}$$

From the derivation on pages 7-8, we know that $(\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1} = (\mathbf{Z}^T\tilde{\mathbf{W}}_2\mathbf{Z})^{-1}$, therefore in theory we have $Var(\hat{\boldsymbol{\beta}}_3) = Var(\hat{\boldsymbol{\beta}}_4)$. However, in statistical packages such as R, it is assumed that we only have the weights correct up to a constant factor $\tilde{\sigma}^2$, i.e., $\mathbf{W}_2 = \tilde{\sigma}^2 Var(\mathbf{Y})^{-1}$ and $\tilde{\mathbf{W}}_2 = \tilde{\sigma}^2 Var(\mathbf{U})^{-1}$ (Stern, 2019). Therefore, $Var(\hat{\boldsymbol{\beta}}_3) = \tilde{\sigma}^2(\mathbf{X}^T\mathbf{W}_2\mathbf{X})^{-1}$, and $Var(\hat{\boldsymbol{\beta}}_4) = \tilde{\sigma}^2(\mathbf{Z}^T\tilde{\mathbf{W}}_2\mathbf{Z})^{-1}$, where $\tilde{\sigma}^2$ in $Var(\hat{\boldsymbol{\beta}}_3)$ is estimated by the weighted mean squared error $WMSE_{individual} = \frac{WSSE_{individual}}{n-p-1}$ for the individual-level analysis; yet in $Var(\hat{\boldsymbol{\beta}}_4)$, $\tilde{\sigma}^2$ is estimated by the weighted mean squared error $WMSE_{group} = \frac{WSSE_{group}}{m-p-1}$ for the group-level analysis. There-

fore,

$$\frac{Var(\hat{\boldsymbol{\beta}}_3)}{Var(\hat{\boldsymbol{\beta}}_4)} = \frac{WSSE_{individual}/(n-p-1)}{WSSE_{group}/(m-p-1)} = \frac{WSSE_{individual}}{WSSE_{group}} \cdot \frac{m-p-1}{n-p-1}$$

$$\Rightarrow\ Var(\hat{\boldsymbol{\beta}}_3) = \frac{WSSE_{individual}}{WSSE_{group}} \cdot \frac{m-p-1}{n-p-1} \cdot Var(\hat{\boldsymbol{\beta}}_4)$$

$$\Rightarrow\ \hat{se}(\hat{\boldsymbol{\beta}}_3) = \sqrt{\frac{WSSE_{individual}}{WSSE_{group}} \cdot \frac{m-p-1}{n-p-1}} \cdot \hat{se}(\hat{\boldsymbol{\beta}}_4)$$

Consider the $WSSE$ for group-level inverse-variance weighted regression analysis:

$$WSSE_{group} = \sum_{j=1}^{m} \frac{n_j \tilde{\varepsilon}_j^2}{\sigma_j^2} = \sum_{j=1}^{m} \frac{n_j (U_j - \hat{U}_j)^2}{\sigma_j^2},$$

where
$\tilde{\varepsilon}_j = U_j - \hat{U}_j$ is the residual of group-level analysis, which can be extracted from the fitted regression models; σ_j^2 is the population variance of birthweight Y_{ij} in group j, which can be extracted from CDC WONDER, and $\frac{\sigma_j^2}{n_j}$ is the variance of the average birthweight U_j for group j.
In contrast, the individual-level inverse-variance weighted regression produces the following $WSSE$

$$\begin{aligned}
WSSE_{individual} =& \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(Y_{ij} - \hat{Y}_{ij})^2}{\sigma_j^2} \\
=& \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(Y_{ij} - U_j + U_j - \hat{Y}_{ij})^2}{\sigma_j^2} \\
=& \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(Y_{ij} - U_j)^2}{\sigma_j^2} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(U_j - \hat{Y}_{ij})^2}{\sigma_j^2} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{2(Y_{ij} - U_j)(U_j - \hat{Y}_{ij})}{\sigma_j^2} \\
=& \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(Y_{ij} - U_j)^2}{\sigma_j^2} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(U_j - \hat{U}_j)^2}{\sigma_j^2} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{2(Y_{ij} - U_j)(U_j - \hat{U}_j)}{\sigma_j^2} \\
=& \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(Y_{ij} - U_j)^2}{\sigma_j^2} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(U_j - \hat{U}_j)^2}{\sigma_j^2} + \sum_{j=1}^{m} \frac{2(U_j - \hat{U}_j)}{\sigma_j^2} \sum_{i=1}^{n_j} (Y_{ij} - U_j) \\
=& \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(Y_{ij} - U_j)^2}{\sigma_j^2} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(U_j - \hat{U}_j)^2}{\sigma_j^2} + \sum_{j=1}^{m} \frac{2(U_j - \hat{U}_j)}{\sigma_j^2} \left(\sum_{i=1}^{n_j} Y_{ij} - n_j U_j \right) \\
=& \sum_{j=1}^{m} \frac{1}{\sigma_j^2} \sum_{i=1}^{n_j} (Y_{ij} - U_j)^2 + \sum_{j=1}^{m} \frac{n_j (U_j - \hat{U}_j)^2}{\sigma_j^2} + \sum_{j=1}^{m} \frac{2(U_j - \hat{U}_j)}{\sigma_j^2} \cdot 0 \\
=& \sum_{j=1}^{m} n_j + \sum_{j=1}^{m} \frac{n_j (U_j - \hat{U}_j)^2}{\sigma_j^2} \\
=& n + WSSE_{group}
\end{aligned}$$

because the population variance of birthweight in group j is $\sigma_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (Y_{ij} - U_j)^2$.

$$\begin{aligned}
\hat{se}(\hat{\boldsymbol{\beta}}_3) &= \sqrt{\frac{WSSE_{individual}}{WSSE_{group}} \cdot \frac{m-p-1}{n-p-1}} \cdot \hat{se}(\hat{\boldsymbol{\beta}}_4) \\
&= \sqrt{\frac{n + WSSE_{group}}{WSSE_{group}} \cdot \frac{m-p-1}{n-p-1}} \cdot \hat{se}(\hat{\boldsymbol{\beta}}_4) \\
&= \sqrt{\frac{(n + WSSE_{group})(m-p-1)}{(n-p-1) WSSE_{group}}} \cdot \hat{se}(\hat{\boldsymbol{\beta}}_4)
\end{aligned}$$

where the standard error $\hat{se}(\hat{\boldsymbol{\beta}}_4)$ can be obtained easily from the output of group-level weighted regression.

We can calculate $WSSE_{group}$ using the formula $WSSE_{group} = \sum_{j=1}^{m} \frac{n_j \bar{\varepsilon}_j^2}{\sigma_j^2}$, or compute it from the group-level weighted mean squared error ($WMSE_{group} = WSSE_{group}/(m-p-1)$) obtained directly from the output of group-level weighted regression in statistical software.

In our PFAS and birthweight study, $n \gg m \gg p$ and $n \gg WSSE_{group}$, thus $\frac{n+WSSE_{group}}{n-p-1} \to 1$, the coefficient $f = \sqrt{\frac{(n+WSSE_{group})(m-p-1)}{(n-p-1)WSSE_{group}}}$ is largely dependent on $\sqrt{\frac{m-p-1}{WSSE_{group}}}$, which depends on the relation between m and $WSSE_{group}$. If $WSSE_{group} \approx m$, then the coefficient $f \approx 1$, the individual-level SE is close to the group-level SE. In the unadjusted model of this study, $WSSE_{group} \approx 100m$, and the coefficient $f \approx 0.1$, indicating that the individual-level SE is around 10% of the group-level SE. In the adjusted and adjusted co-exposure models in this study, $WSSE_{group} \approx 2m$, and the coefficient $f \approx 0.7$, so the individual-level SE is around 70% of the group-level SE.

Using the percentage of water measurements with PFAS detection as the exposure metric, the standard errors in the crude models were reduced by 91-92%, and the standard errors in the adjusted and adjusted co-exposure models were reduced by 32% in inverse-variance weighted individual-level analysis compared with the inverse-variance weighted group-level analysis.

Using the population-weighted average PFAS concentration as the exposure metric, the standard errors in the crude models were reduced by 92%, and the standard errors in the adjusted and adjusted co-exposure models were reduced by 34-35% in inverse-variance weighted individual-level analysis compared with the inverse-variance weighted group-level analysis.

In summary, although the standard errors were reduced substantially by variance correction from group-level analysis to individual-level analysis, our conclusions were the same as those in the paper.

```
### R function to correct the confidence interval for regression parameter
### in  individual-level inverse-variance weighted regression
f <- function(data = data, mod = mod){
# input multiple-stratified data and group-level inverse-variance weighted regression model
require(broom)
n <- sum(data$Births)
# sum up the number of births in each group to get the total number of births
m <- nrow(data)
# number of groups
wsse.group <- sum(data$Births*(mod$residuals^2)/data$SD^2, na.rm = T)
# weighted SSE for group-level inverse-variance weighted regression
p <- nrow(tidy(mod)) - 1
# number of regression parameters except for the intercept
frac <- sqrt( (n+wsse.group)*(m-p-1)/((n-p-1)*wsse.group) )
# see the above formula: the coefficient of the standard error of
# regression parameter for group-level analysis to get the standard error of
# regression parameter for individual-level inverse-variance weighted analysis
lo <- tidy(mod)$estimate[2] - frac * tidy(mod)$std.error[2] * qnorm(0.975)
# lower limit of confidence interval of regression parameter
# for individual-level inverse-variance weighted analysis
hi <- tidy(mod)$estimate[2] + frac * tidy(mod)$std.error[2] * qnorm(0.975)
# upper limit of confidence interval of regression parameter
# for individual-level inverse-variance weighted analysis
print(c(lo, hi))
# print the confidence interval of regression coefficient
# for individual-level inverse-variance weighted analysis
print(1-frac)
# percentage of reduction in standard error compared with
# the group-level inverse-variance weighted analysis
}
```

References:

Hal Stern. Statistical Notes of Statistics 210, University of California, Irvine. 2019. https://www.ics.uci.edu/ sternh/
Ramsey FL, Schafer DW. The Statistical Sleuth: A Course in Methods of Data Analysis. Third Edition. Cengage Learning, 2012.