



# THE EFFECTS OF SAMPLE DESIGN ON STATISTICAL INFERENCE FROM THE UNIVERSITY OF MICHIGAN DIOXIN EXPOSURE STUDY

**Biling Hong, James Lepkowski, Kristen Olson, Elizabeth Hedgeman, Qixuan Chen, Shih-Yuan Lee, Chiung-Wen Chang, Barbara Lohr-Ward, Kathleen Ladronka, Brenda Gillespie, Alfred Franzblau, Peter Adriaens, Avery Demond, David Garabrant**  
University of Michigan, Ann Arbor, Michigan, USA

## INTRODUCTION & OBJECTIVES

- Linear regression models were fit to identify factors that explain variation in **serum TEQ concentration** (based on 2005 TEFs) measured from 946 participants in the University of Michigan Dioxin Exposure Study (UMDES). The participants were sampled from five geographically-defined populations in Midland, Saginaw, part of Bay, Jackson and Calhoun Counties in Michigan, using a two-stage area probability household sample design.
- The regression analyses accounted for sampling weights reflecting selection probabilities and non-response propensities, stratification, clustering, and imputation variance to ensure the inferences from the regression models were applicable to the population from which participants were selected. Statistical results that do not adjust for sample design often generate incorrect inferences. **The goal of this poster is to address the effects of the complex sample design used in our study.**

## METHODS

- Outcome variable  $\log_{10}$  (serum TEQ concentration)** was regressed on household dust, soil and all potential predictors derived from the UMDES questionnaire (for example, basic demographic and health variables, residential history, property use, work history, recreational activities in the contaminated areas, food consumptions including meat, fish, game meat, etc.) to identify significant predictors by applying a multi-stage backward selection. All statistical analyses used **SAS** version 9.1.
- Missing values in all predictors were imputed by using a sequential regression multiple imputation procedure in **IVeware**. All the presented results were based on 5 imputations.
- Three linear regressions were performed with the same outcome and predictors, but treating the 946 multiply imputed cases in UMDES in three different ways:
  - Simple Random Sample (SRS) without adjusted for sampling weights;
  - SRS adjusted for sampling weights;
  - Complex sample survey data, adjusted for sampling weights, stratification and clustering.

## RESULTS

**Table 1** shows the distributions of sampling weights for 946 serum samples by region. In Midland/Saginaw floodplain and near floodplain, sampling weights were relatively small (about 7 to 9) compared to areas out of floodplain and Jackson/Calhoun counties (about 344 to 374). It reflected the unequal sample probabilities across the regions.

**Table 1: Descriptive Statistics of sampling weights for serum samples in UMDES study**

Region	N	Mean	Median	75th percentile	Std. Dev.
Overall	946	181	67	284	253
M/S FP	243	7	6	9	5
M/S Near FP	205	9	7	9	9
M/S Out FP	204	374	342	458	283
M/S Plume	43	126	77	118	143
Jackson/Calhoun	251	344	271	439	249

Note: M/S: Midland and Saginaw; FP: Floodplain.

**Table 2: The differences in parameter estimate and the standard error of the estimate for most important predictors across three different linear regressions.**

Important predictors	1. SRS_unweighted		2. SRS_weighted		3. Complex survey data			
	Estimate	S.E.	Estimate	S.E.	Deff <sup>1</sup>	Estimate	S.E.	Deff <sup>2</sup>
age*	0.0162	0.0022	0.0203	0.0019	0.74	0.0203	0.0026	1.97
age <sup>2</sup> *	-0.0001	1.9E-05	-0.0001	1.7E-05	0.82	-0.0001	2.4E-05	1.91
BMI*	0.0073	0.0017	0.0073	0.0015	0.74	0.0073	0.0025	2.74
BMI loss in the past 12 months	0.0097	0.0032	0.0106	0.0027	0.72	0.0106	0.0028	1.07
Gender (1 for female, and 0 for male)*	0.0985	0.0715	0.1033	0.0609	0.73	0.1033	0.0859	1.99
Num of months the 1st child was breast-fed*	-0.0048	0.0019	-0.0051	0.0015	0.59	-0.0051	0.0015	1.00
Num of months for all children except 1st one were breast-fed	-0.0022	0.0009	-0.0018	0.0006	0.46	-0.0018	0.0007	1.39
Pack-years of smoking*	-0.0016	0.0003	-0.0021	0.0003	0.86	-0.0021	0.0005	3.19
Living in M/S FP vs. living in J/C*	-0.0135	0.0193	-0.0337	0.0467	5.85	-0.0337	0.0276	0.35
Living in M/S Near FP vs. living in J/C*	-0.0208	0.0188	-0.0051	0.0451	5.76	-0.0051	0.0270	0.36
Living in M/S out FP vs. living in J/C*	-0.0174	0.0188	-0.0251	0.0147	0.61	-0.0251	0.0216	2.16
Living in M/S Plume vs. living in J/C*	-0.0605	0.0284	-0.0374	0.0315	1.24	-0.0374	0.0281	0.79
Num. of years lived in M/S in 1960-1979	0.0033	0.0009	0.0029	0.0009	1.10	0.0029	0.0014	2.42
Soil dioxin concentration in house perimeter 0-1**	-0.0001	4.6E-05	-0.0002	0.0001	6.90	-0.0002	0.0001	1.20
Soil dioxin concentration in garden*	4.2E-05	0.0001	0.0005	0.0003	18.92	0.0005	0.0004	2.24
The highest soil concentration found in each property	2.3E-05	1.1E-05	9.9E-06	1.0E-05	0.96	9.9E-06	3.8E-06	0.14
Household dust dioxin loading $\mu\text{g}/\text{m}^2$	-1.6E-05	1.4E-05	-2.3E-05	1.3E-05	0.84	-2.3E-05	1.1E-05	0.77
Did water activities near the Tittabawassee R. in 1960-1979 ( $\geq 1$ per month vs. never)	0.04	0.0381	0.25	0.0742	3.80	0.2516	0.0961	1.68
Did fishing activities in the Saginaw R. or Bay after 1980 ( $\geq 1$ per month vs. never)	0.1143	0.0280	0.0959	0.0294	1.10	0.0959	0.0297	1.02
Did hunting activities in the surrounding areas of the Saginaw R. or Bay in 1960-1979 (Y vs. N)	0.0790	0.0306	0.1200	0.0375	1.50	0.1200	0.0474	1.60
Did hunting activities in the surrounding areas of the Saginaw R. or Bay after 1980 ( $\geq 1$ per month vs. never)	-0.04	0.0724	-0.22	0.1214	2.81	-0.2244	0.0685	0.32

- The results of predictors related to property use, work history, and fish and game meat consumptions are not presented here, since they only contributed a small percentage of the variance in the serum TEQ concentration.
- \* Variable was forced into the regression model.
- Deff<sup>1</sup>: Variance under SRS\_weighted / Variance under SRS\_unweighted
- Deff<sup>2</sup>: Variance under complex sample design / Variance under SRS\_weighted

**Table 2 presents the results of three different regressions:**

- The Deffs (see definition at bottom of Table 2) of most predictors ranged from 0.5 to 3 with an average of 2.
- Pack-years of smoking** had relatively large design effect (Deff<sup>2</sup>=3.19) compared with other demographics predictors. The variance estimate increased by about 3 times after adjustment for the complex sample design.
- Living in M/S FP vs. living in J/C** showed strong design effects: (1) Deff<sup>1</sup>=5.85 indicated the variance estimate had increased by about 6 times due to sampling weights; (2) Deff<sup>2</sup>=0.35 indicated that precision was gained by using regions as sample strata. **Living in near FP vs. living in J/C** had similar results.
- Deff<sup>1</sup> for **soil TEQ in garden and in house perimeter top 1 inch** was 18.92 and 6.9, respectively, probably due to the strong correlation within geographic areas of these values.
- Deff<sup>2</sup> for the **highest soil TEQ found in each property** was very small (0.14), indicating that precision was gained by using regions as sample strata.
- The sampling weights itself had large impact on the parameter and variance estimates of the **frequency of water activities near the Tittabawassee River** (parameter changed from 0.04 to 0.25, and variance increased by about 4 times). **The frequency of hunting activities near the Saginaw River or Bay** had a similar impact from sampling weights on the estimate, but the variance decreased 68% after taking the complex sample design into account.

## CONCLUSIONS

- Although complex sample design often leads to increased variances, the analyses based on 946 cases in the UMDES show that both increased and decreased variance can occur.
- Predictors that correlated with sampling weights were more likely to increase the variance, such as soil dioxin concentration in the garden and house perimeter soil in the top 1 inch. Predictors varied across regions decreased variance by using region as a sampling stratum, such as the frequency of hunting activities near the Saginaw River or Bay.

Financial support for this study comes from the Dow Chemical Company through an unrestricted grant to the University of Michigan.