

VARIABLE SELECTION METHODS INFLUENCE THE IDENTIFICATION OF FACTORS THAT PREDICT SERUM DIOXIN CONCENTRATIONS IN MICHIGAN, USA

Biling Hong¹, David Garabrant¹, Qixuan Chen¹, Chiung-Wen Chang¹, Xiaohui Jiang¹, Elizabeth Hedgeman¹, Brenda Gillespie¹, James Lepkowski¹, Alfred Franzblau¹
¹University of Michigan, Ann Arbor, Michigan, USA;

Introduction and Objectives

- Linear regression models were performed to identify important factors that were associated with serum dioxin concentrations in the Midland and Saginaw Counties in Michigan, using data from 946 participants in the University of Michigan Dioxin Exposure Study (UMDES) that were selected from the study area by a complex sample design.
- We used two different variable selection approaches in the linear regression models: backward selection and forward stepwise selection. The influential diagnostics were then performed to investigate the influence of influential observations on the regression coefficients, by using DFBETAS.
- The purpose of this paper is to compare the results from these approaches and investigate which approach is more sensitive to the influential observations in the data.

Methods

- In linear regression when the number of potential predictors is relative large compared with the number of observations, there are typically not enough degrees of freedom to run a single step of backward selection from the complete variable list. Although this problem is often addressed by using a multi-stage **backwards selection procedure (Figure 1)**, there is no guarantee that the resulting model will be optimal.
- Forward stepwise selection (Figure 2)** allows variable selection in a single step, but there is no software available to implement this in the setting of multiply-imputed survey data. We solved this problem by writing a SAS macro that implements forward stepwise selection in the setting of multiply-imputed survey data.
- To assess the effect of an individual observation on each estimated parameter of the fitted model, the **DFBETAS diagnostic** (the standardized difference in the parameter estimate due to deleting the observation) was calculated for each observation.

Figure 1: Backward Selection Strategy

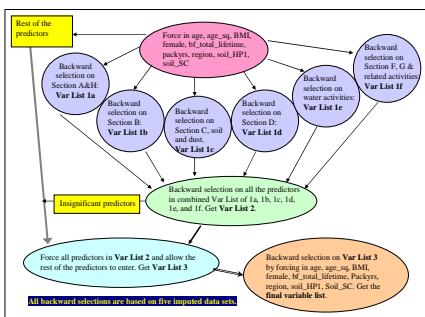
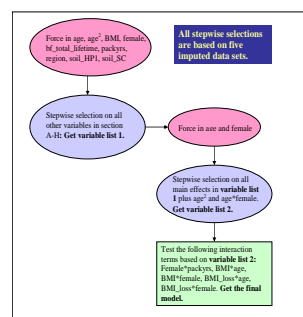


Figure 2: Stepwise Selection Strategy



Results

Table 1: Number of total and unstable predictors in the models.

Note: Explanatory factors that were statistically significant ($p < 0.05$) when all observations were included, but which became non-significant when three or fewer observations were omitted, were felt to be **unstable** since their inclusion was dependent on, in many cases, a single influential observation.

Num. of predictors in the model	TEQ		2378-TCDD	
	Backward	Stepwise	Backward	Stepwise
Total	27	18	30	24
Unstable	11	5	13	7

Table 2: Model result for TEQ (Backward vs. Stepwise)

Predictors	Backward Selection		Stepwise Selection	
	Estimate	Pvalue	Estimate	Pvalue
TEQ				
age_center	0.0112	0.0030	0.0107	0.0030
age_sq_center	-0.0001	0.0017	-0.0001	0.0050
BMI_center	0.0073	0.0040	0.0066	0.0179
bf_total_lifetime	-0.0029	0.0000	-0.0030	0.0000
female	0.0380	0.0288	Yes	0.0291
packys	-0.0021	0.0000	-0.0020	0.0002
BMI female center	-0.0082	0.0048	-0.0076	0.0139
BMI loss	0.0130	0.0000	0.0106	0.0071
female age center	0.0035	0.0001	0.0033	0.0004
preg_nochildren			0.0117	0.0131
midland_60_79	0.0028	0.0049	0.0029	0.0060
clip_40_59	0.0061	0.0024	0.0069	0.0022
firedamage_40_59	-0.0897	0.0147	Yes	
firedamage_60_79	0.0310	0.0000	Yes	0.0296
weedkiller_60_79	-0.0049	0.0010	Yes	
Soil TEQ max*	7.99E-06	0.0041	Yes	
Dust TEQ loading*	-1.98E-05	0.0468	Yes	
clow_40_59			0.0123	0.0000
emerg_resp_after80	-0.0065	0.0065	Yes	
herbicide_weed_60_79	0.0119	0.0182	Yes	
other_work_60_79	-0.0079	0.0235	Yes	
water_otherriver_60_79_c_1	-0.0381	0.0057	Yes	
water_otherriver_60_79_c_2	-0.0067	0.6783		
water_littliver_60_79_c_1	0.2584	0.0114	Yes	0.1708
water_littliver_60_79_c_2	-0.0303	0.4964		-0.0505
F4_after80_vrs	0.0033	0.0018	0.0031	0.0033
otherfish_sag_rb_c_1	-0.2708	0.0000	-0.2541	0.0000
otherfish_sag_rb_c_2	0.0136	0.6300	-0.0179	0.5029
fishing_sag_rb_after80_c_1	0.0824	0.0031	0.0830	0.0014
fishing_sag_rb_after80_c_2	0.0271	0.1868	0.0269	0.1831
gmeat_tpgqw_skin	0.0580	0.0322	Yes	
hunt_sag_rb_60_79_d	0.1146	0.0132	Yes	
hunt_sag_rb_after80_c_1	-0.1977	0.0023	Yes	
hunt_sag_rb_after80_c_2	-0.0883	0.0587		

* Soil TEQ concentration around the house was identified in backward selection and statistically significantly associated with increased serum TEQ (p-value=0.0041). If the most influential observation is excluded from the regression analysis, there is not a statistically significant association between the soil TEQ concentration and serum TEQ. But this factor was not identified by the forward stepwise selection even when the most influential observation was included in the regression analysis.

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

Results, cont.

Table 3: Model result for 2378-TCDD (Backward vs. Stepwise)

Predictors	Backward Selection		Stepwise Selection	
	Estimate	Pvalue	Estimate	Pvalue
2378-TCDD				
BMI_loss_age_center	0.0101	0.0000	0.0097	0.0000
BMI_center	0.0047	0.0506		
bf_total_lifetime	-0.0032	0.0178	-0.0047	0.0000
female	0.0941	0.0007	0.1222	0.0000
packys	-0.0032	0.0046	-0.0031	0.0074
Sardin soil 2378-TCDD	0.0069	0.0005	Yes	0.0052
BMI_loss	0.0167	0.0052	Yes	0.0188
industrial_d	-0.0981	0.0074	Yes	-0.0907
female_age_center	0.0098	0.0000	Yes	0.0082
white	-0.1180	0.0032	Yes	-0.1467
midland_40_59*	0.0048	0.0370		
midland_60_79	0.0095	0.0000	0.0115	0.0000
firedamage_60_79	0.0229	0.0145	Yes	0.0234
lostburned_40_59	-0.0084	0.0000	Yes	0.0084
water_enter_littliver	0.1074	0.0481	Yes	0.1074
clow_40_59			0.0221	0.0000
flow_after80			0.0071	0.0453
emerg_resp_40_59			0.0643	0.0000
emerg_resp_60_79	0.0141	0.0007	Yes	
emerg_resp_after80	-0.0189	0.0000	Yes	-0.0215
herbicide_weed_40_59	0.0144	0.0030	Yes	0.0128
million_wetman_vrs	0.0064	0.0000	Yes	0.0064
waste_scrap_water_60_79	0.0148	0.0042	Yes	0.0486
water_sagrbvrs_after80_c_1	-0.2715	0.0093	Yes	-0.2917
water_sagrbvrs_after80_c_2	-0.1203	0.0115	Yes	-0.0488
water_littliver_after80_c_1	0.2968	0.0006	Yes	0.2917
water_littliver_after80_c_2	0.0805	0.0093	Yes	0.0488
F10_after80_vrs	0.0072	0.0088		
F4_after80_vrs	0.0060	0.0002	0.0049	0.0027
wp_sag_rb_c_1	-0.2448	0.0005	Yes	
wp_sag_rb_c_2	-0.1489	0.0089	Yes	
otherfish_sag_rb_c_1			-0.1750	0.0000
otherfish_sag_rb_c_2			0.0090	0.0000
fishing_sag_rb_after80_c_1	0.1888	0.0003	Yes	
fishing_sag_rb_after80_c_2	0.0651	0.1587	Yes	
ES_60_79_vrs	-0.0078	0.0189	Yes	
hunt_60_79_c_1	0.3473	0.0002	Yes	0.3643
hunt_60_79_c_2	-0.0068	0.9539		0.0160
hunt_litr_after80_c_1	-0.5411	0.0000	Yes	-0.5692
hunt_litr_after80_c_2	-0.0550	0.3604	Yes	-0.0289
sag_dairy_ellse_mtlpw_d	-0.1688	0.0007	Yes	

* Living in Midland or Saginaw Counties in the 1940s and 1950s was identified in backward selection and was statistically significantly associated with increased serum 2378-TCDD (p-value=0.037). If the most influential observation is excluded from the regression analysis, there is not a statistically significant association between the years living in Midland or Saginaw Counties in the 1940 and 1950s and serum 2378-TCDD. But this factor was not identified by the forward stepwise selection even when the most influential observation was included in the regression analysis.

Conclusions

- The most important predictors of serum dioxins (age, gender, body mass index, smoking status, the length of breastfeeding, etc.) were consistently identified by using either backward or forward stepwise variable selection.
- However, some factors that were dependent on a small number (1 to 3) of observations tended to be identified in backward selection, but not in forward stepwise selection. These factors should be interpreted with caution insofar as the associations were highly dependent on a few influential observations.
- Based on these findings, we recommend using forward stepwise variable selection in linear regression analysis when the number of potential explanatory factors is large.