Investigating the Effects of DDE and PCB Exposure on Premature Delivery

Youngsoo Baek, Yunran Chen and Xiaojun Zheng

Executive Summary

We study the association between exposure to Dichlorodiphenyldichloroethylene (DDE) and 12 Polychlorinated Biophenyl (PCB) members and the risk of premature delivery. To address the collinearity caused by PCB members, we consider both a logit regression model, either adjusted for a weighted sum of standardized PCB levels, or for a few PCB variables selected through LASSO regression. Both models suggest higher exposure to both DDE and PCB is associated with higher odds of preterm delivery.

Introduction

Dichlorodiphenyltrichloroethane (DDT) and Polychlorinated biphenyl (PCB) were proven to be toxic, leading to a ban on use in United States in 1972 and 1978, respectively. We are interested in whether and how the two chemicals relate to premature delivery risk. World Health Organization (WHO) defines preterm births as babies born alive before 37 weeks of pregnancy are completed. Different delivery times correspond to subsantially different levels of risk for infants' health. World Health Organization mentions three subcategories of preterm birth: extremely pre-term (<28 weeks), very preterm (28 to 32 weeks), and moderate to late preterm (32 to 37 weeks).

The dataset considered is a subsample of the US Collaborative Perinatal Project (CPP) conducted by Longecker et al. (2001), which contains concentration doses for DDE and PCBs as well as the gestational ages of 2,380 women in 12 medical centers. Other observed features include triglycerides and cholesterol, which are relevant since the chemicals are lipophylic; smoking status; and demographic characteristics, including race, age and socio-economic index. Due to unavailability of controlled experimental study, we aim to use generalized linear regression models to adjust for as many predictors as possible and infer whether there exists significant association between exposure to DDE/PCB and increased preterm delivery risks.

Materials and Methods

In order to model the linear association between the probability of preterm delivery and predictors of interest on some suitable scale, we consider a logistic regression model, where the log-odds of premature delivery are modeled by a linear combination of DDE and PCB levels. Binary logit model, a standard classification method, suffers loss of information on different delivery periods. Therefore, we bin the gestational age to reflect increasing levels of risk. In particular, the response are binned into: very preterm (27 to 32 weeks); moderately preterm (32 to 33 weeks); late preterm (34 to 36 weeks); and normal deliveries (37 to 46 weeks).

The ordered logit model, where categories are denoted by $k = 1, 2, 3, 4$, ordered from normal to very preterm, can be expressed as follows:

$$\Pr(Y_i \leq k|\mathbf{x}_i) = \text{logit}^{-1}(\alpha_k + \mathbf{x}_i^T\beta) \iff \text{Odds}(Y_i \leq k|\mathbf{x}_i) = e^{\alpha_k}e^{\mathbf{x}_i^T\beta}.$$

The important assumption of this model, which extends the binary logit model, is that the odds for different categories are proportional by a constant determined by fixed effects. The data exhibit some indication that the model assumption can be violated, an issue we revisit in Section 5.

For any regression model estimation, high correlation between 12 different PCB chemicals pose challenge (see Figure 2). High correlation between predictors inflates standard errors of the estimated effects, so our inference for all predictors in the model can be severely limited. Two approaches are discussed. First, we exclude all PCB levels and includ the first principal component from the principal components analysis (PCA). Since this principal component is a weighted sum of all PCB members that are standardized to unit scale, it can be thought of as a proxy variable for "adjusted total" PCB levels. Direct interpretation of the effect magnitude based on regression estimates, however, is not feasible. Alternatively, we fit LASSO regression, a regression placing penalty on inclusion of more variables, to exclude a number of PCB variables. The selected variables are then included in the logistic model, so the effect estimates for individual PCB variables are interpretable.

Results

Exploratory Data Analysis

Predictors adjusted for in the model include DDE (μg), PCB-related variables (originally measured in ng), triglycerides and cholesterol levels (g/dL), race, maternal age, and smoking status. The three score variables corresponding to the subjects' income, occupation, and education were not recorded for about 20% of the entire sample. In a PCA-logistic model, these variables were excluded based on the results of F-test against model not including score variables. In a LASSO-logit model, they were also considered for variable selection after imputation. Sensitivity analysis result is shown in Section 3.3.

Figure 3 shows heterogeneity in racial composition of subjects across different centers in the study. Categorical variables for each subject, indicating which center she has delivered at, were thus added in the regression model. These variables are shifts in the mean preterm delivery risk corresponding to each center. There are alternative approaches to modeling this heterogeneity than this "fixed effect" model, as we discuss at the end.

Main Results

Based on the binary logistic regression model, the significant predictors include DDE, proxy variable for PCB levels, triglyceride and cholesterol levels, along with mean shifts in centers 15, 37, and 82. These centers comprise primarily of black race subjects; accordingly, we see insignificant effect for race variable. Estimated model effects and their 95% confidence intervals can be found in Table 1. Most of the estimates have a direct scientific interpretation. For instance, for a 1μg increase in DDE exposure, adjusted for all other variables, a mother has 0.8% increased odds of premature delivery. Similarly, for a 1g/dL increase in triglyceride level, adjusted for other variables, a mother has 0.3% increased odds of premature delivery. Such inference, however, is hampered for PCB, as discussed in Section 2. Ordinal logistic model estimates mostly agree in the signs and magnitudes of the effects; however, cholesterol level is no longer considered a significant variable.

LASSO logit model that minimizes the binomial deviance criterion estimates non-zero coefficients for PCBs 074 and 153. Likelihood ratio test against the model excluding these two variables

suggest evidence of non-zero, significant effect these two chemicals in particular have on preterm delivery risks (Table 3). The coefficient estimates and confidence intervals of the sequentially re-fit linear model, of which the signs and magnitudes are consistent with the estimates of the previous approach, can be found in Table 2. Overall, our models agree that there is evidence of significant effect of DDE and PCB exposures on increased risk of pre-term deliveries.

The LASSO logit model slightly improves the Bayesian information criterion (BIC) of the PCA-logistic model (1625/2054 for binary/ordered, versus 2061/2643). Lower values of BIC indicate better explanatory power of the model, adjusted for penalties to including too many predictors.The former approach has additional advantage of retaining the interpretability of selected PCB effects on data scale: a 1ng increase in PCB 074 exposure, adjusted for other included variables, corresponds to 44.5% increased odds of premature delivery.

Sensitivity Analysis

LASSO model considers all observed features as possible variables to be selected. Predictive mean matching algorithm was used to impute the missing entries before fitting the model. For both binary and ordered LASSO-logit fits, sensitivity analyses were performed by comparing the estimated log odds with and without imputed values (see Figs. 5 and 6). We observe specifically the interval estimates for DDE and PCB variables are quite robust to imputed values.

Discussion

All of our models are limited by the fact that the estimates and standard errors overstate our certainty one way or another, as the additional uncertainty caused by modeling decision for multiple PCB variables is not accounted for. Alternative methods we have explored but not included in this report include Bayesian estimation methods with different models. In particular, a hierarchical regression that incorporates heterogeneity bettween centers as random effects will be desirable over fixed effects model. Furthermore, the Bayesian factor model that models different PCB variables by a latent factor can overcome the previously discussed limitations of the PCA-logit model, including lack of direct interpretation of the magnitude of the effect. Fitting such complex models, however, can confront numerical instabilities in estimation unless we carefully parametrize the model.

Another natural extension of our model will be allowing each coefficient β to vary by different delivery periods, or even binned preterm delivery categories. Such model overcomes the proportional odds assumption discussed in Section 2, which is both restrictive and difficult to diagnose through the residuals. Finally, modeling interactions will be a natural extension that may call for Bayesian methods with a strong prior given our relatively small sample size.

Tables and Figures

EDA

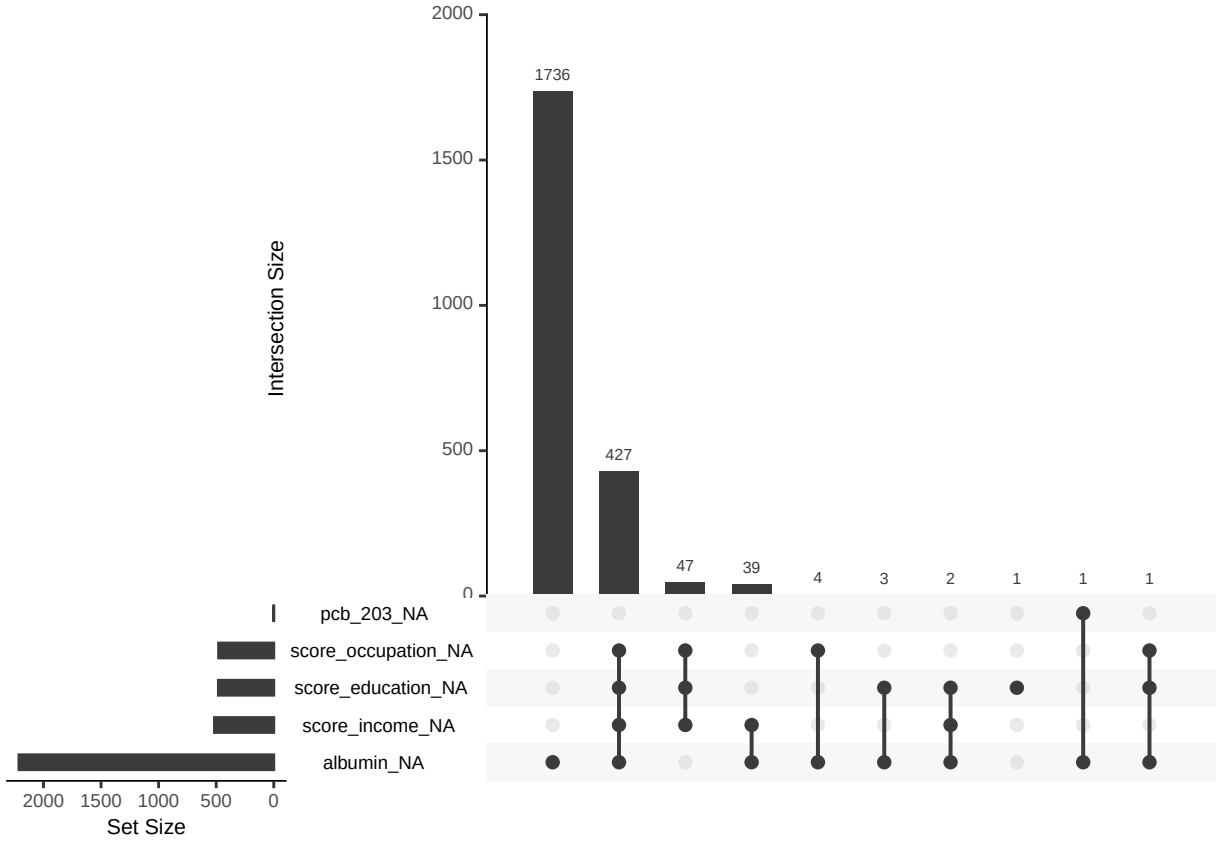

Figure 1: Missing Data

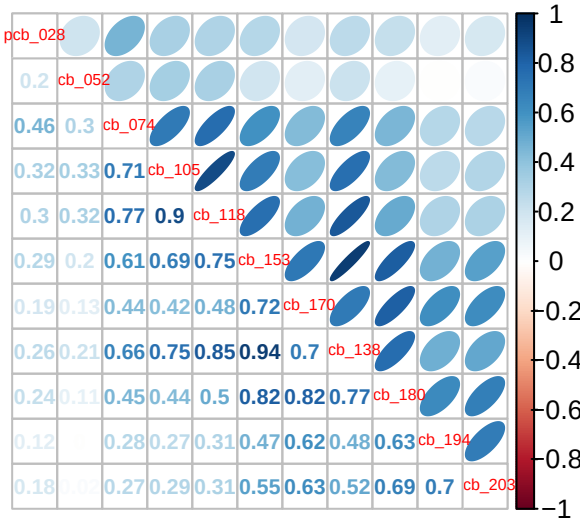

Figure 2: Correlation plot across PCBs

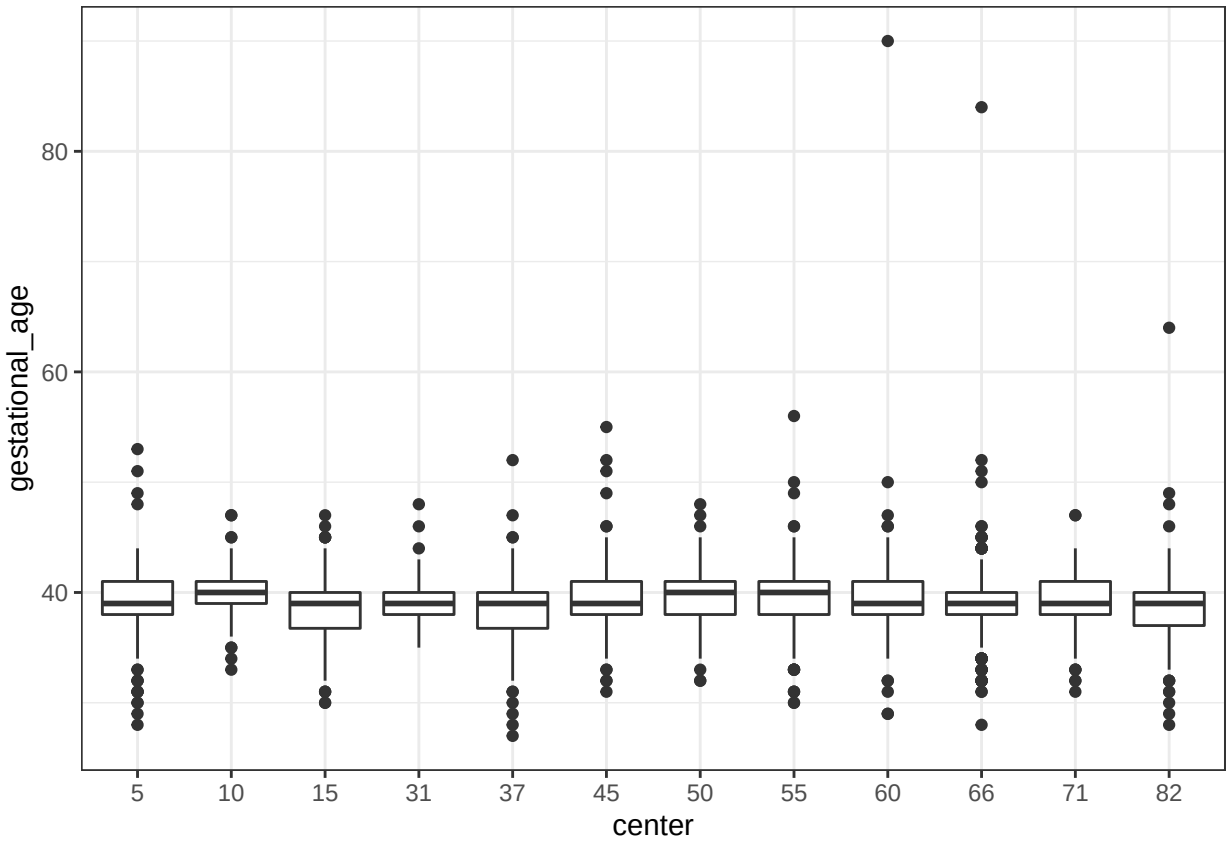

Figure 3: Heterogeneity across Centers

PCA-logistic

	Lower bound	Mean	Upper bound
Very preterm	0.003	0.008	0.018
Moderately	0.010	0.022	0.049
Late preterm	0.040	0.086	0.174

Lasso-logistic

Table 1: Logistic

	Mean	2.5 %	97.5 %
(Intercept)	-2.412	-3.212	-1.619
dde	0.009	0.003	0.014
PC1	0.076	0.021	0.130
triglycerides	0.003	0.002	0.005
raceblack	0.227	-0.183	0.640
raceother	0.460	-0.238	1.136
maternal_age	-0.011	-0.030	0.009
smoking_status	0.183	-0.054	0.420
cholesterol	-0.003	-0.005	-0.001
center10	-1.041	-2.121	-0.183
center15	1.022	0.419	1.628
center31	-0.536	-1.574	0.343
center37	0.879	0.371	1.387
center45	0.291	-0.308	0.875
center50	0.009	-0.688	0.642
center55	0.550	-0.131	1.218
center60	0.466	-0.161	1.059
center66	0.368	-0.142	0.882
center71	0.072	-0.548	0.649
center82	0.660	0.053	1.268

Table 2: Ordinal logistic

	Mean	2.5 %	97.5 %
dde	0.008	0.002	0.014
PC1	0.081	0.026	0.134
triglycerides	0.003	0.001	0.004
raceblack	0.232	-0.176	0.642
raceother	0.466	-0.236	1.145
maternal_age	-0.010	-0.030	0.009
smoking_status	0.171	-0.064	0.407
cholesterol	-0.002	-0.004	-0.001
center10	-1.037	-2.116	-0.179
center15	1.054	0.454	1.657
center31	-0.544	-1.581	0.332
center37	0.864	0.359	1.368
center45	0.277	-0.319	0.856
center50	0.013	-0.684	0.645
center55	0.582	-0.101	1.252
center60	0.490	-0.137	1.081
center66	0.369	-0.137	0.879
center71	0.089	-0.529	0.664
center82	0.674	0.070	1.278

Table 3: Testing for PCBs in logisitic and Ordinal logistic

Deviance	Df	Pr(>Chi)
8.595	2	0.014
10.429	2	0.005

Table 4: Logistic

	Mean	2.5 %	97.5 %
(Intercept)	-2.457	-3.481	-1.447
dde	0.008	0.001	0.015
pcb_074	0.426	-0.350	1.160
pcb_153	0.368	-0.081	0.811
triglycerides	0.003	0.001	0.005
raceblack	0.392	-0.103	0.890
raceother	0.828	-0.188	1.770
maternal_age	-0.009	-0.031	0.013
smoking_status	0.170	-0.107	0.446
cholesterol	-0.002	-0.004	0.000
center10	-0.880	-1.985	0.015
center15	0.601	-0.130	1.334
center31	-1.003	-2.330	0.080
center37	0.690	0.091	1.291
center45	0.022	-0.694	0.724
center50	0.022	-0.749	0.720
center55	-0.316	-1.555	0.804
center60	0.264	-0.488	0.966
center66	0.031	-0.592	0.659
center71	-0.139	-0.899	0.560
center82	0.427	-0.302	1.156
score_education	-0.004	-0.010	0.002
score_income	-0.002	-0.008	0.003
score_occupation	-0.003	-0.009	0.002

Table 5: Ordinal logistic

	Mean	2.5 %	97.5 %
dde	0.007	0.001	0.014
pcb_074	0.585	-0.164	1.285
pcb_153	0.340	-0.098	0.771
triglycerides	0.003	0.001	0.004
raceblack	0.413	-0.081	0.909
raceother	0.918	-0.114	1.871
maternal_age	-0.008	-0.030	0.014
smoking_status	0.163	-0.111	0.437
cholesterol	-0.002	-0.004	0.000
center10	-0.885	-1.992	0.012
center15	0.634	-0.093	1.363
center31	-1.012	-2.342	0.070
center37	0.718	0.120	1.318
center45	0.033	-0.679	0.731
center50	0.020	-0.750	0.716
center55	-0.325	-1.582	0.813
center60	0.315	-0.435	1.014
center66	0.046	-0.572	0.671
center71	-0.101	-0.858	0.595
center82	0.463	-0.261	1.188
score_education	-0.004	-0.010	0.003
score_income	-0.003	-0.008	0.003
score_occupation	-0.003	-0.009	0.002

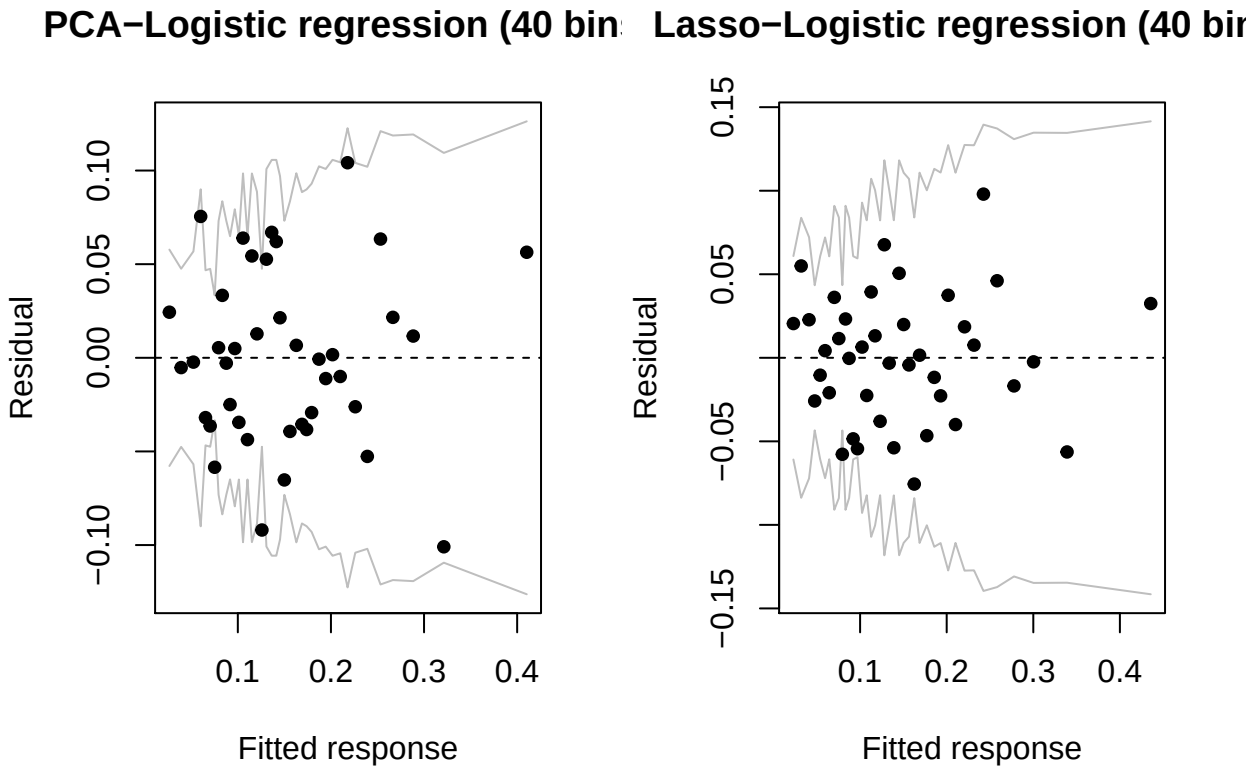

Figure 4: Diagnosis of Lasso-logistic

Sensitivity Analysis

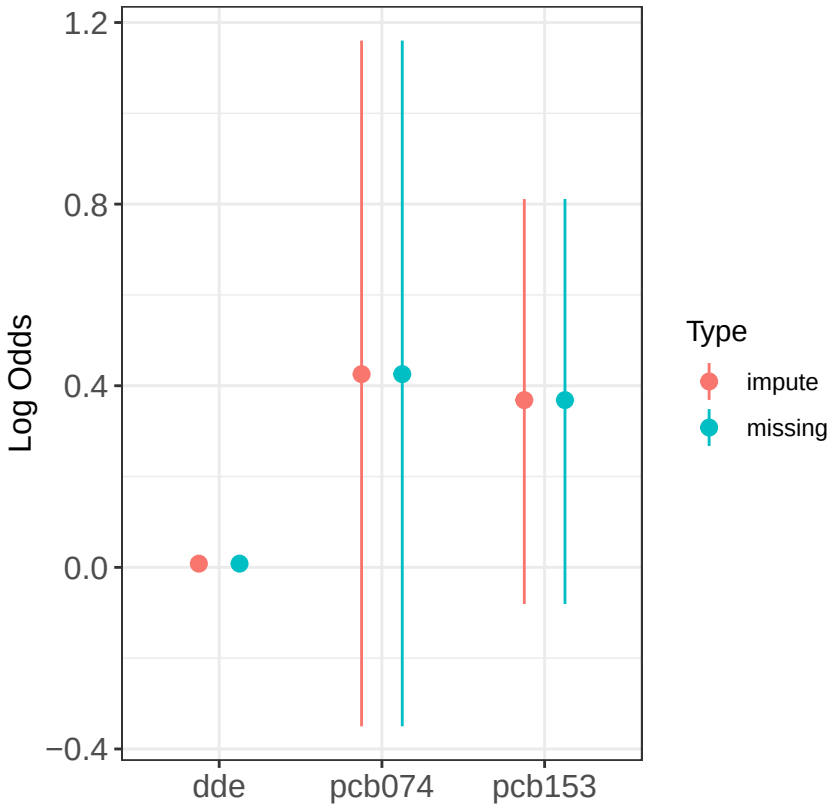

Figure 5: Sensitivity Analysis (Binary)

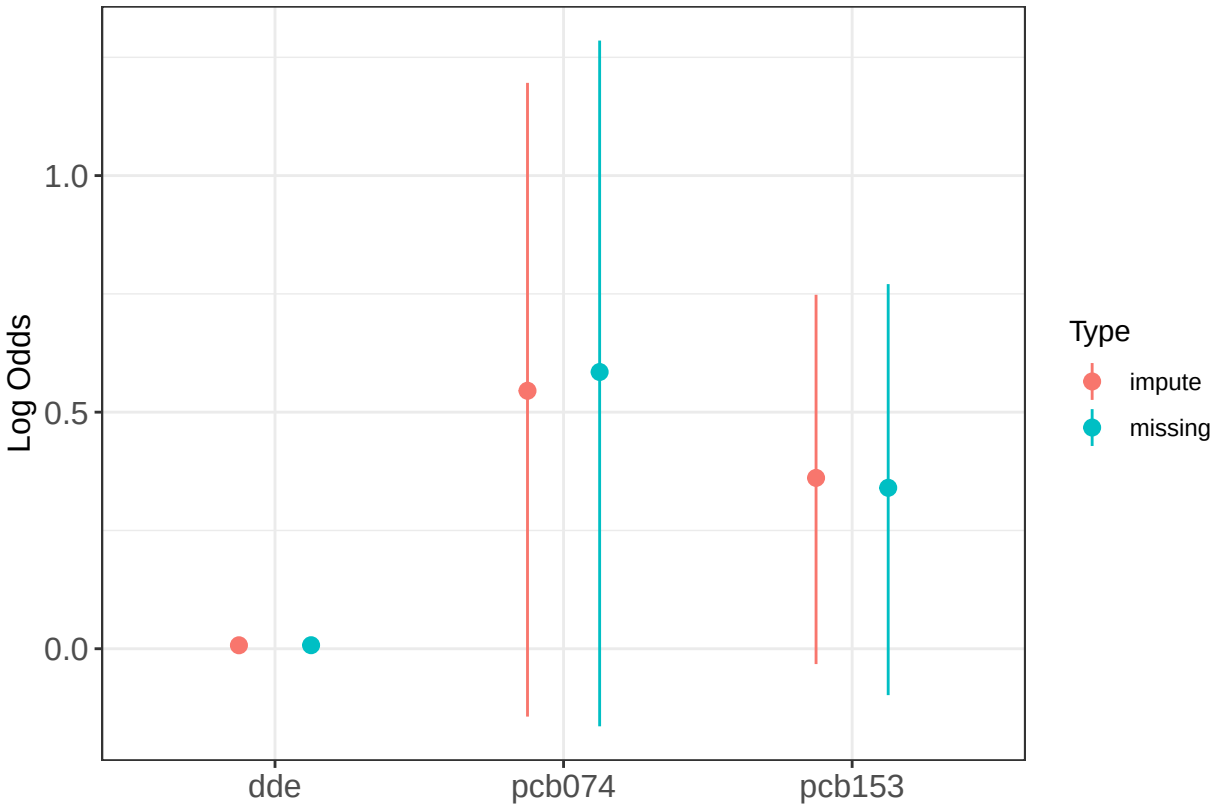

Figure 6: Sensitivity Analysis (Ordinal),fig.width=4.5