

Assessing Effects of Exposures to DDE and PCBs on Premature Delivery via Ordinal Logistic Regression

Raphael Morsomme and Rihui Ou and Alessandro Zito

1. Introduction

Dichlorodiphenyldichloroethylene (DDE) and Polychlorinated Biphenyls (PCB) are two chemical elements which were commonly in use in the United States and were banned during the 70's due to their detrimental effect on human health. Exposure to these chemical products has been linked to XXX. Because they remain in the environment and in the human body for a long period of time, studying their impact on human health is particularly important.

In this study, we examine the effect of DDE and PCB on fetus; more precisely, we attempt to estimate the impact of exposure to these chemicals elements to early delivery. To accomplish this, we use an ordinal logistic regression model. We find that the impact of these chemical components is race specific: exposure to DDE increases the risk of early birth among white people and exposure to PCB increases the risk among non white people.

Section 2 introduces the data and Section 3 presents our methodology. Section 4 describes our finding. We discuss these and conclude the report in Section 5 concludes.

2. Data

The data set consists of 2,380 pregnant women that visited a hospital during their pregnancy in 2001 and contains the length of the gestation in weeks and the concentration doses of DDE and 12 PCB breakdown products in the blood. It also contains the concentration of cholesterol and triglycerides in the women's blood along with demographic information (race, level of education, income, occupation, age, smoking status and the center attend by the woman).

mention that PCB and DDE are stored in fatty tissues

3. Method

3.1. Feature Engineering

First, since we focus on the impact of DDE and PCB on early delivery, we dichotomize the dependent variable *gestational duration* measured in weeks into three levels: at risk for gestation shorter than 33 weeks, pre term for gestation between 34 and 36 week, and at term for gestation longer than 37 weeks. Second, we also aggregate the 12 PCB measurement into one variable since these variables are highly correlated (see Figure 4). To accomplish this, we take their average after standardizing them in order to prevent one measurement

to dominate the aggregate variable. Third, we combine the measurements of chemical in blood with the fat-related variables to estimate the level chemical to which the women were exposed. We start by estimating the total amount of fatty tissues using the following formula ([reference](#))

$$\text{lipid} = 2.27 * \text{cholesterol} + \text{triglycerides} + 0.623. \quad (1)$$

Since the amount of chemical absorbed is proportional to the amount of fatty tissue one has, we divide the concentration of PCB and DDE in blood by the estimated amount of fatty tissue in order to obtain an estimate of the initial level of exposure to these chemical product, that is,

$$\text{DDE}_{\text{exposure}} = \frac{\text{DDE}}{\text{lipid}} \quad (2)$$

and

$$\text{PCB}_{\text{exposure}} = \frac{\text{PCB}_{\text{aggregate}}}{\text{lipid}}. \quad (3)$$

Finally, we combine *black* ($n = ??$) and *other* ($n = ??$) in the variable *race* in order to have levels with a sufficient amount of data.

3.2. Ordinal Logistic Regression Model

$\text{DDE}_{\text{exposure}}$ and $\text{PCB}_{\text{exposure}}$ are the variables of interest. In order to identify non-spurious association between these variable and the occurrence of early deliveries, we control for 7 background attributes: [list them. interaction with centers and race to allow different effect across centers and race](#)

Since not all attributes are necessarily related to the dependent variable, we conduct an backward AIC-based variable selection procedure. Ideally, we would have preferred to avoid selecting variables and instead run a BMA, but since Merlise's package does not allow ordinal logistic regression, we opted for the simpler frequentist approach. The variable selection procedure includes [list variables](#) and drops [list variables](#).

We use the variables selected y the procedure as predictors in a Bayesian ordinal logistic regression model. The model was fit using stan. We use 10 chains of 10,000 iterations each, and put a uniform prior on the coefficients.

3.3. Model Checking

Since the ordinal data is used here, the common residual plot model checking is not applicable here. Instead, the surrogate residual method suggested by ([??](#)) is used. If the model assumption is satisfied, the surrogate residual R_S should display three features:

1. $E(R_S|X) = 0$
2. $\text{Var}(R_S|X) = c$, the conditional variance of R_S is constant
3. The empirical distribution of R_S resembles an explicit distribution that is related to the link function $G^{-1}(\cdot)$. Specifically, $R_S \sim G(c + \int u dG(u))$.

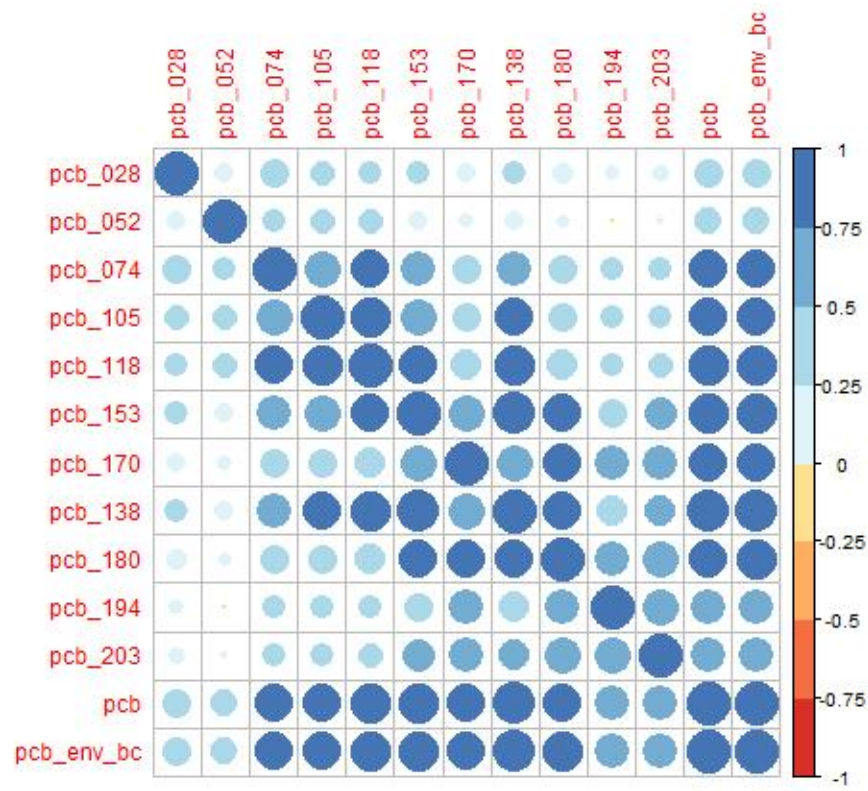


Figure 1: Correlation among the 12 PCBs variables.

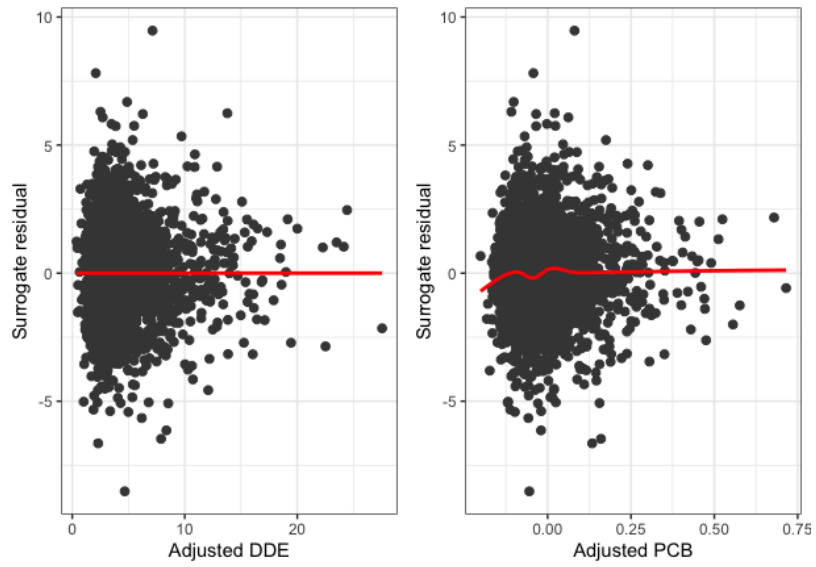


Figure 2: Surrogate residuals of DDE and PCB

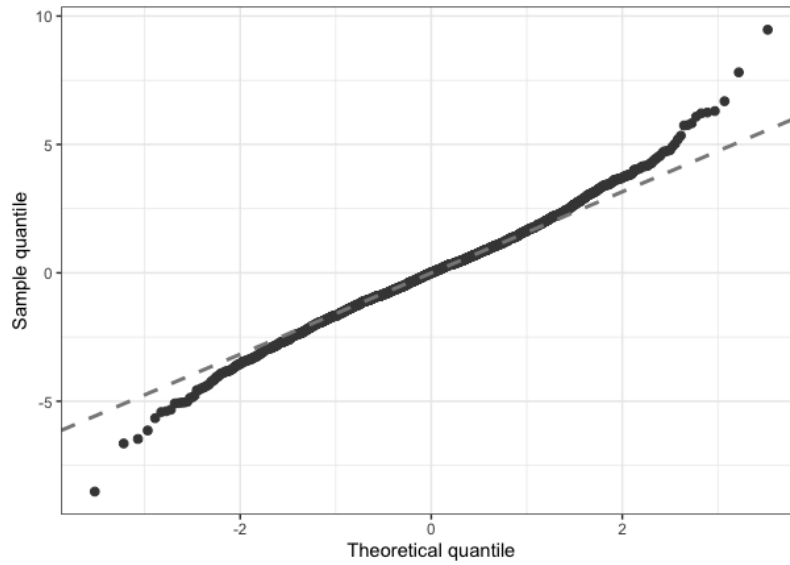


Figure 3: QQ plot of the Surrogate residuals

The scatterplot (Figure 2) indicates that feature 1 and 2 are roughly satisfied. The QQ plot indicates that feature 3 is roughly satisfied, although the tail of our sample distribution is lighter than that of the theoretical one.

Table 1: Example of student data

Student ID	Course ID	Academic Year	Period	Grade
44940	CAP3000	2009-2010	4	8.8
37490	SSC2037	2009-2010	4	8.4
71216	HUM1003	2010-2011	4	6.8
44212	SSC2049	2010-2011	2	8.4
85930	SSC2043	2011-2012	1	4.3
14492	COR1004	2012-2013	2	8.5
34750	HUM2049	2013-2014	5	6.0
32316	SSC1001	2013-2014	1	8.5
22092	SCI1009	2014-2015	1	6.4
19512	COR1004	2016-2017	5	7.0

	mean	5%	95%
DDE _{exposure}	0.02	-0.01	0.05
PCB _{exposure}	1.76	0.72	2.75
DDE _{exposure} *white	0.05	-0.02	0.12
PCB _{exposure} *white	-1.60	-3.26	0.02

Table 2: 90% credible intervals and posterior mean of coefficients

4. Results

We gave the 90% credible intervals for coefficients and their posterior mean. (Table 2). The interpretation of coefficients are as follow:

- DDE_{exposure}: For a 1 unit increase of DDE_{exposure}, holding other covariates constant,
 - **Nonwhite**: the odds of having a more dangerous delivery increase by $(e^{(0.02)} - 1) * 100\% = 2.02\%$. The 90% credible interval is $[-2.00\%, 4.12\%]$.
 - **White**: the same odds increase by $(e^{(0.02+0.05)} - 1) * 100\% = 7.25\%$
- PCB_{exposure}: For a 0.1 unit increase of PCB_{exposure}, holding other covariates constant,
 - **Nonwhite**: the odds of having a more dangerous delivery increase by $(e^{(1.76)*0.1} - 1) * 100\% = 19.22\%$. The 90% credible interval is $[6.47\%, 30.65\%]$.
 - **White**: the same odds increase by $(e^{0.1*(1.76-1.60)} - 1) * 100\% = 1.595\%$

Main findings: the effect of the chemicals on the risk of early delivery is race dependent. Exposure to DDE has a particularly detrimental impact on the gestation process among white women, while exposure to PCB affects non-white more (see Figure ??).

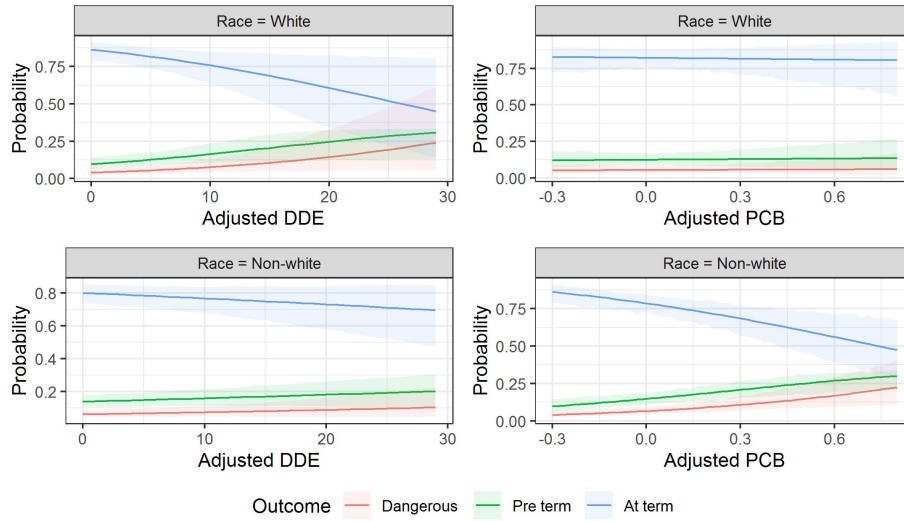


Figure 4: Estimated probability of gestation outcomes in function of race, and exposure to DDE and PCB.

5. Conclusion

Future, add quadratic term to age since gestations at a young and an old age are more at risk of complications.

Consider interaction between PCB and DDE