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 for agricultural purpouses and were banned during the 70's due to their detrimental effect on human health. Exposure to these chemical products has been linked to neurobehavioral and developmental deficits in newborns. As the human body stores them in its fatty tissues, studying their impact on human health is particularly important.

In this report, we examine the effect of DDE and PCB on fetuses; more precisely, we assess the potential association between the exposure to these chemicals and the chance of early delivery. Ideally, a higher exposure to the substances induces a preterm delivery, which may have adverse consequences for the child. To verify this theory, we construct an ordinal logistic regression model over three delivery groups We find that the impact of the substances is essentially race specific: exposure to DDE increases the risk of early childbirth among white people and exposure to PCB increases the risk among non-white ones.

The report is divided as follows: Section 2 presents the data and our methodology, Section 3 reports our findings, Section 4 discusses the results and concludes.

2. Method

2.1. Data

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

The data were not clean. In particular, 43 women showed a length of gestation superior to 45 weeks (the second longest gestation period ever recorded), and 1 women did not show any records of the PCBs levels. Thus, we decide to drop them, reducing the observations to 2,336. Finally, we mean impute the missing data on the income, education and occupational scores[1]

1. Note that these scores will not end up in the final model.

2.2. 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)

$$\text{lipid} = 2.27 * \text{cholesterol} + \text{triglycerides} + 0.623. \tag{1}$$

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,

$$\text{DDE}_{\text{exposure}} = \frac{\text{DDE}}{\text{lipid}} \tag{2}$$

and

$$\text{PCB}_{\text{exposure}} = \frac{\text{PCB}_{\text{aggregate}}}{\text{lipid}}. \tag{3}$$

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

2.3. 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.

2.4. 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:

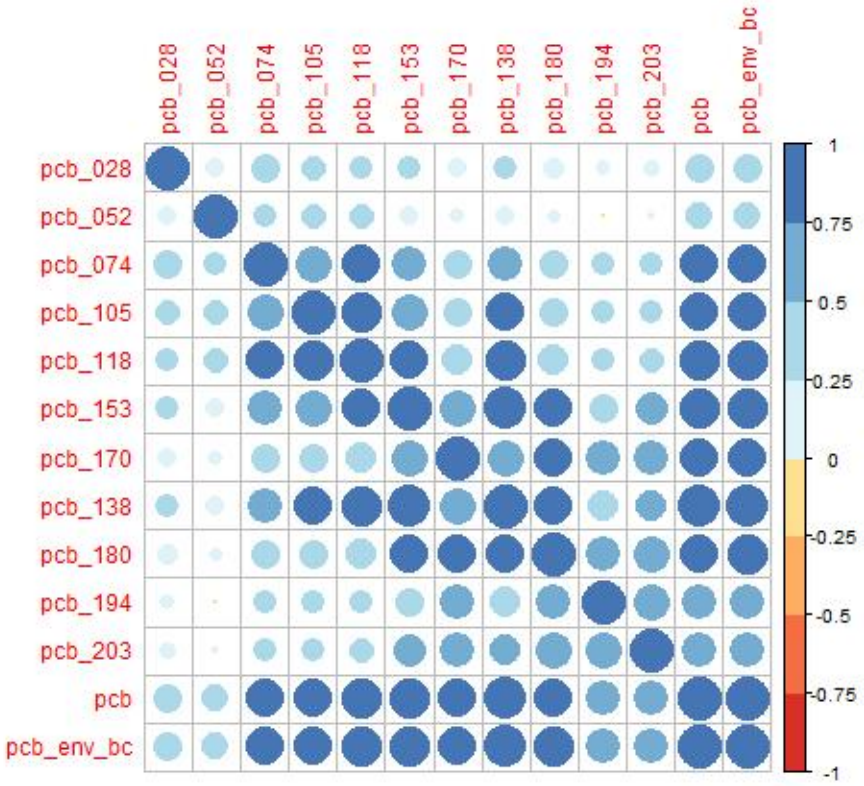

Figure 1: Correlation among the 12 PCBs variables.

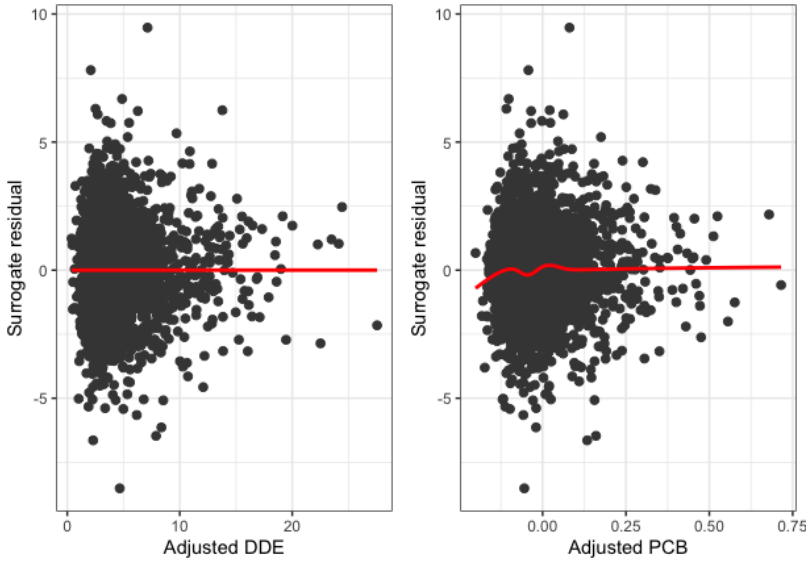

Figure 2: Surrogate residuals of DDE and PCB

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

The scatterplot (Figure2) 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

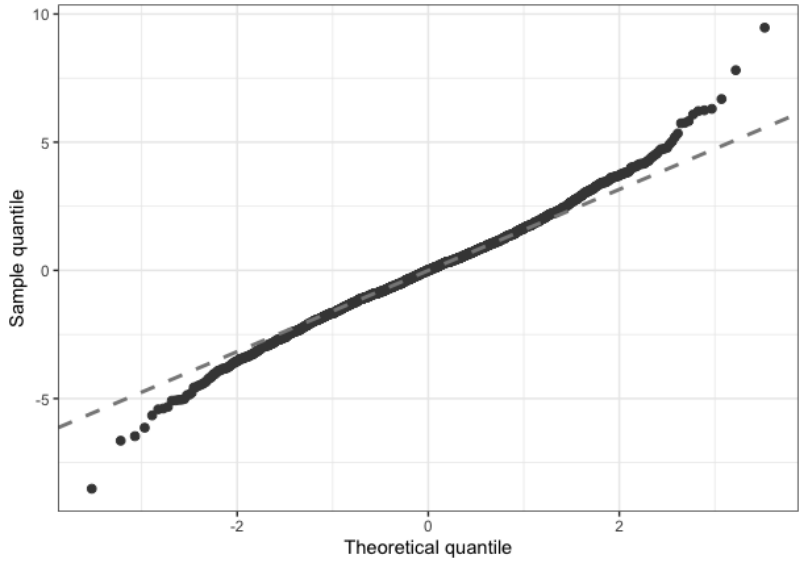

Figure 3: QQ plot of the Surrogate residuals

3. Results

3.1. EDA

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

- $\text{DDE}_{\text{exposure}}$: For a 1 unit increase of $\text{DDE}_{\text{exposure}}$, 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 $(e^{(0.02+0.05)} - 1) * 100\% = 7.25\%$
- $\text{PCB}_{\text{exposure}}$: For a 0.1 unit increase of $\text{PCB}_{\text{exposure}}$, 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 $(e^{0.1*(1.76-1.60)} - 1) * 100\% = 1.595\%$

3.2. Main Findings

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 **??**)).

	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

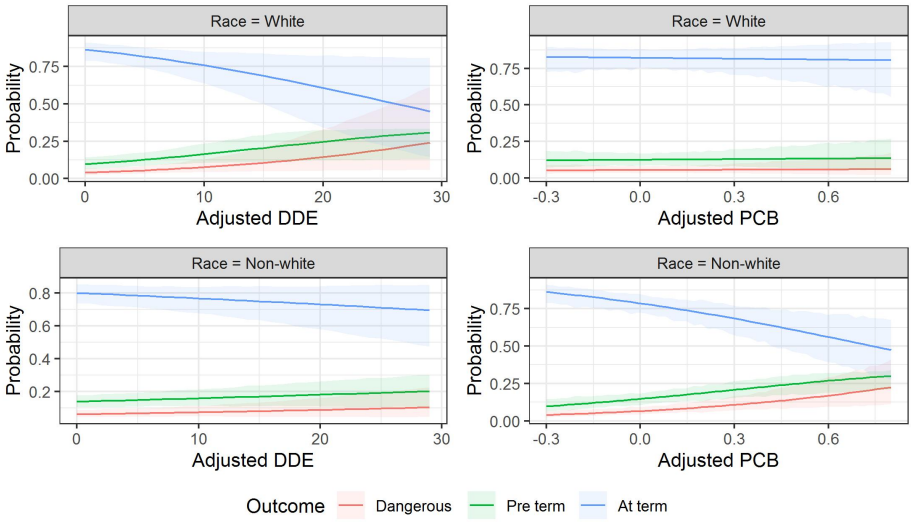

Figure 4: Estimated probability of gestation outcomes in function of race, and exposure to DDE and PCB.

3.3. Sensitivity Analysis

4. 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