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

Raphael Morsomme and **Rihui Ou** and **Alessandro Zito**

Abstract

1. Introduction

Dichlorodiphenyldichloroethylene (DDE) and Polychlorinated Biphenyls (PCB) are two chemical elements which were commonly useD in the United States for agricultural purpouses and were banned during the 70's due to their detrimental effect on human health. In particular, 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 defined by the recorded lenght of the gestation period. 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. Methods

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). However, 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 number of women to 2,336. Finally, we mean impute the missing data on the income, education and occupational scores[1]. The following subsection describes the construction of the relevant variable in our analysis.

2.2. Feature Engineering

As a first step, we divide the women into three groups, labelled as "Dangerous Preterm", "Preterm" and "At term" based upon the lenght of thei gestation (shorter than 33 weeks, between 34 and 36 weeks and longer than 37 weeks, respectively). The groups are mean to capture the danger associated to the birth for the child himself. While a delivery after 37 weeks is considered normal, the main organs (especially the respiratory system) develop between week 34 and 37, making a birth before 34 weeks more dangerous. Second, as the twelve PCB measurements showed a high correlation (see Figure 4), we aggregate them into a unique variable by taking their standardized average [2]. Third, we combine the

1. Note that these scores will not end up in the final model.
2. We first standardize each single PCB to to prevent one measurement to dominate the aggregate variable.

measurements of chemical in blood with the fat-related variables to estimate the initial level of DDE and PCBs to which the women were exposed from the environment. In particular, we calculate the total amount of fatty tissues using the formula in PHILLIPS (1989) BERNET (2007)

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

Then, 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 log of the level of lipid[3].

$$\text{DDE}_{\text{exposure}} = \frac{\text{DDE}}{\log(\text{lipid})} \qquad \text{PCB}_{\text{exposure}} = \frac{\text{PCB}_{\text{aggregate}}}{\log(\text{lipid})}. \tag{2}$$

Finally, we aggregate the women into two groups, "white" and "non-white", based on the reported race[4].

2.3. Ordinal Logistic Regression Model

$\text{DDE}_{\text{exposure}}$ and $\text{PCB}_{\text{exposure}}$ are our variables of interest, whereas the above constructed delivery group is our dependent variable. In order to identify non-spurious association between these variables and the occurrence of early deliveries, we run the following ordinal logistic regression model

$$\text{logit}(P(gestgroup \leq j)) = \beta_{0j} - \mathbf{X}\boldsymbol{\beta} \tag{3}$$

In order to incorporate our model uncertainty, we wish to adopt a Bayesian model averaging (BMA). Yet, since we are not aware of existing computing to realize BMA on an ordinal logistic regression model, we first selected variables using a forward and backward AIC-based procedure. The procedures selects XXX and drops YYY. We then use the selected variables as predictors in a Bayesian ordinal logistic regression model. The model was fit on stan and we 10 chains of $10,000$ iterations each were run.

Sensitivity Analysis

In order to conduct a sensitivity analysis, we run the Bayesian model using two different priors: uniform and R^2. The results are consistent under the two priors (see refer tables)

3. Results

3.1. EDA

Figure 4 presents the correlation matrix of the 12 PCB measurements. We can observe that they are highly correlated with each other. This is not surprising since they correspond to breakdown products of PCB. This provides a rationale for aggregating these measurements into one variable that attempts to approximate the total amount of PCB in the women's

3. This correction derives from a Box-Cox analysis of our model, following the basic procedure in (LI LONGNECKER DUNSON (2013)). See the appendix for further details.
4. The original data had 1,016 white women, 1,201 black ones, and 120 labelled as "other". As the categories are unbalanced, we prefer to merge for a clearer interpretation.

	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 1: 90% credible intervals and posterior mean of coefficients

blood (see Section 2.2). Figure 3 presents the distribution of gestational outcome across the different centers. There exists a wide spread of outcome among the centers: for instance, while center 31 did not have any delivery that was *dangerous*, more than a third of the deliveries realized in the centers $15, 62, 37$ occurred pre-term. Figure 3 shows the distribution of the estimated exposure to PCB and DDE per gestational outcome and per race. We can observe a negative weak association between gestational outcome and DDE. We also note that the distribution of the chemicals vary across the two race levels. This indicates that we need to control for race in the regression model in order to prevent the variable from acting like a confounder.

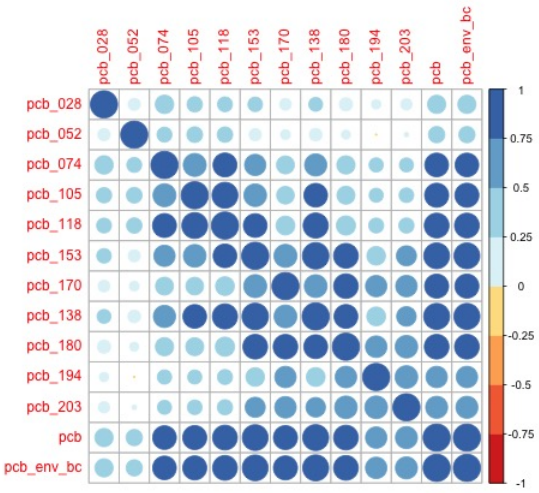

Figure 1: Correlation among the 12 PCBs variables.

3.2. Main Findings

ADD HERE THE SAME PLOT UNDER A R2 PRIOR.

RIHUI: MAKING A LISTED DOT LIKE WE DID IN THE SLIDES IS NOT A GOOD IDEA WHEN YOU HAVE TO SUMMARIZE EVERYTHING IN THREE PAGES. THIS SECTION NEEDS TO BE REVISED. Main findings: the effect of the chemicals on the

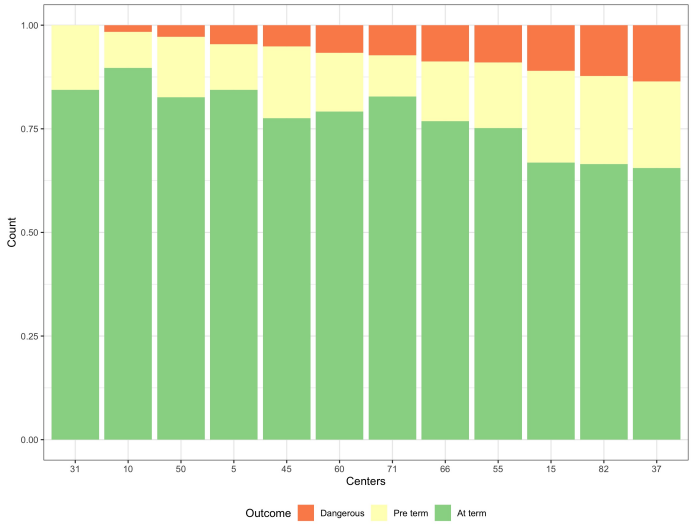

Figure 2: Gestational outcome per hospital center.

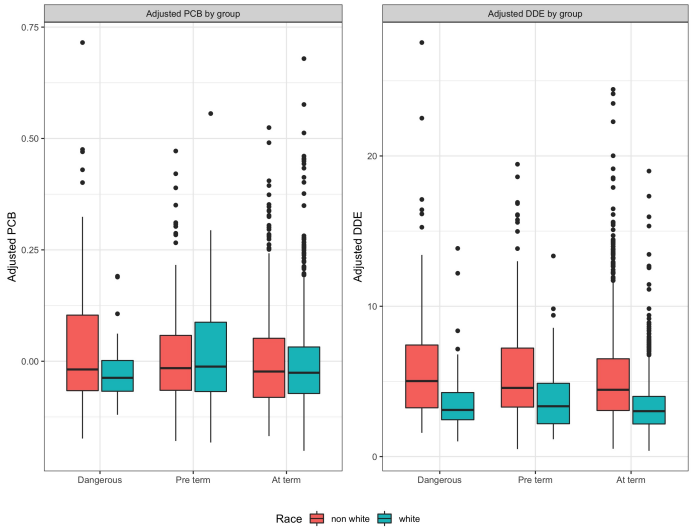

Figure 3: Distribution of estimated exposure to PCB and DDE per gestational outcome and per race.

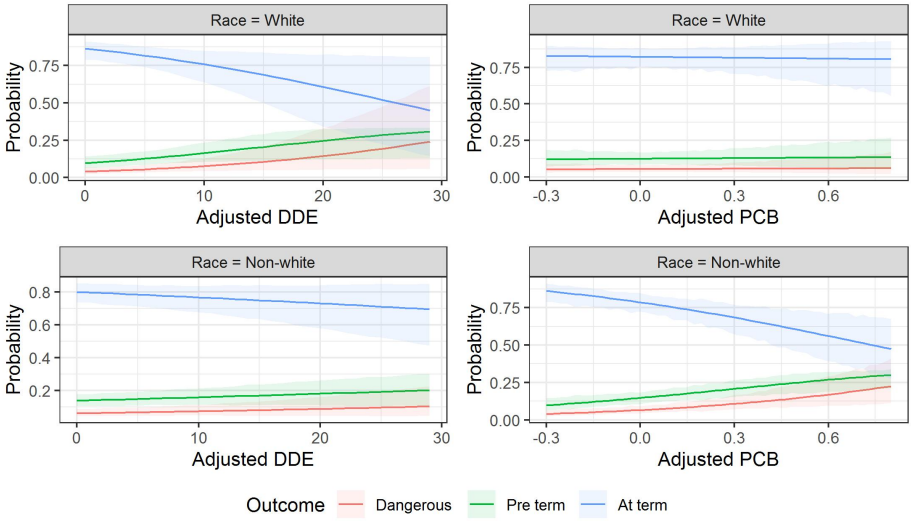

Figure 4: Estimated probability of gestation outcomes in function of race, and exposure to DDE and PCB.

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 **??**)). We gave the 90% credible intervals for coefficients and their posterior mean. (Table 1). 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 $(e^{0.1*(1.76-1.60)} - 1) * 100\% = 1.595\%$

3.3. Sensitivity Analysis

RIHUI: WE NEED TO RUN THE BAYESIAN MODEL WITH A DIFFERENT PRIOR. AND COMMENT ON THE RESULTS HERE.
RAPAHEL: SHOULD WE INCLUDE THE FREQUENTIST ESTIMATES ASWELL?

4. Conclusions and further discussion

Future directions: 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 (chemicals

commonly interact with each other), model the effect of the chemicals in a non-linear way since small levels of exposure are likely to have no effect on human health and we expect the effect tha stabilize past a certain threshold.

Appendix A. Appendix

A.1. Box-Cox analysis for lipid adjustment.

ALESSANDRO'S JOB

A.2. Variable selection procedure

RIHUI'S JOB. WE NEED TO SHOW THE RESULT OF OUR AIC CODE. MORE IN GENERAL, WE NEED TO SHOW SOME FREQUENTIST RESULTS.

A.3. Model Checking

RIHUI'S JOB. FIX IT AND MAKE IT BETTER READABLE. WHAT ARE THE SURROGATE RESIDUALS? WE NEED AN EXPLANATION. THIS IS NOT SUFFICIENT.

Since the ordinal data is used, the common residual plot model checking is not applicable here. Instead, the surrogate residual method suggested by () is used. The surrogate residual is defined as $R_S = S - E(S|X)$, where S is some continuous variable generated from the conditional distribution of latent variables given Y. If the model assumption is satisfied, the surrogate residual R_S should display three characteristics:

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, $R_S \sim G(c + \int u dG(u))$, where c is a constant.

The scatterplot (Figure5) 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.

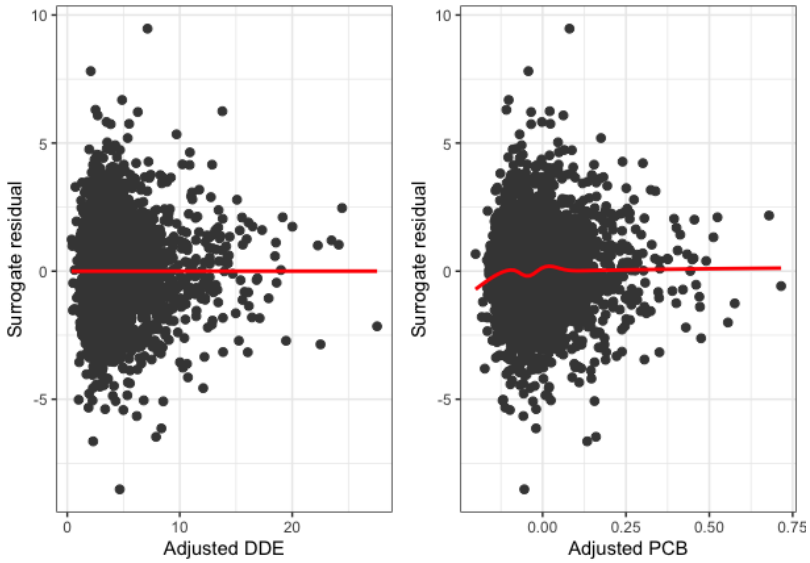

Figure 5: Surrogate residuals of DDE and PCB

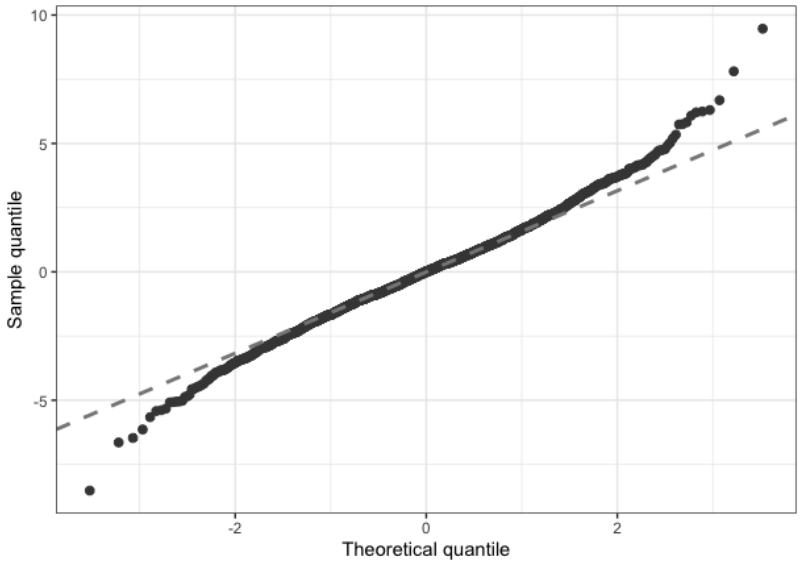

Figure 6: QQ plot of the Surrogate residuals