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

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

Abstract

Li et al (2013)

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 1 in Appendix A), we aggregate them into a unique variable by taking their standardized average [2]. Third, we combine the 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 et al. (1989) and Bernert et al (2007)

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

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.

Then, since the amount of chemical absorbed is proportional to the amount of fatty tissue one has, we adjust the concentration of PCB and DDE in blood by dividing 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

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

for $j = 0$ (=Dangerous), 1 (=Preterm), where β_{j0} is the baseline coefficient for category j, $\boldsymbol{\beta}$ is the effect of each covariate on the log odds and $\mathbf{X}$ is the matrix of regressors. In order to incorporate the model uncertainty into the analysis, we adopt a Bayesian approach. In particular, we first run a forward and backward AIC-based variable selection, which leads to an $\mathbf{X}$ that includes $\text{DDE}_{\text{exposure}}$, $\text{PCB}_{\text{exposure}}$, smoke, center, race and the interactions between ($\text{DDE}_{\text{exposure}}$ and race and between $\text{PCB}_{\text{exposure}}$ and race.[5]. Second, we set a uniform prior over each coefficent. To further check the robustness of the model, we apply a R^2-prior as well.The followign section reports and summarises our findings.

3. Results

3.1. EDA

Figure 1 presents the correlation matrix of the 11 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 blood (see Section 2.2). Figure 2 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 does not have any *dangerous* delivery, more than a third of the deliveries recorded in 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 groups. This indicates that we need to control for race in the regression model in order to prevent the variable from acting as a confounder.

3. This correction derives from a Box-Cox analysis of our model, following the basic procedure in Li et al (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.
5. Ideally, we wish to adopt a Bayesian model averaging (BMA) approach. However, as we are not aware of existing computing resources to apply BMA on an ordinal logistic regression model, we carry out the selection using a frequentist procedure. The variables dropped from the full model are three scores, mother age, and the interactions between $\text{DDE}_{\text{exposure}}$ and center, and between $\text{PCB}_{\text{exposure}}$ and center.

3.2. Main Findings

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. We fit our Bayesian model via `Stan` using 10 chains of $10,000$ iterations each. 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 4)). 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 $(e^{0.1*(1.76-1.60)} - 1) * 100\% = 1.595\%$

3.3. Sensitivity Analysis

We vary different priors (prior on R^2 with location 0.3, 0.5 and 0.8 respectively) to check if the model is sensitive to priors (See Table A). We attempted to check with other priors, like Cauchy, but it's impossible to specify such prior in the *stan_polr* function of this package. To conclude, our method is not sensitive to the choice of priors.

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.

mice. Conducting variable selection on on mean-imputed data reduces the chance of the imputed variables to be selected. Our selection procedure drops the three variable that are mean imputed.

References

Li, D; Longnecker, M.P.; and Dunson, D.B.
Lipid Adjustment for Chemical Exposures: Accounting for Concomitant Variables.
Epidemiology, Nov 2013

Phillips, D; Pirke, J., Burse, V.; Bernert, J.; Henderson, L.; Needham, L.
Chlorinated hydrocarbon levels in human serum: Effects of fasting and feeding.
Archives of Environmental Contamination and Toxicology, 1989

Bernert, JT.; Turner, WE.;, Patterson, DG. Jr;, Needham, LL.
Calculation of serum total lipid concentrations for the adjustment of persistent organohalogen toxicant measurements in human samples.
Chemosphere, 2007

Liu, D.; and Zhang, H.;
Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach
Journal of the Americal Statistical Association, 2018

Appendix A. Figures and Tables

Figure 1: Correlation among the 12 PCBs variables.

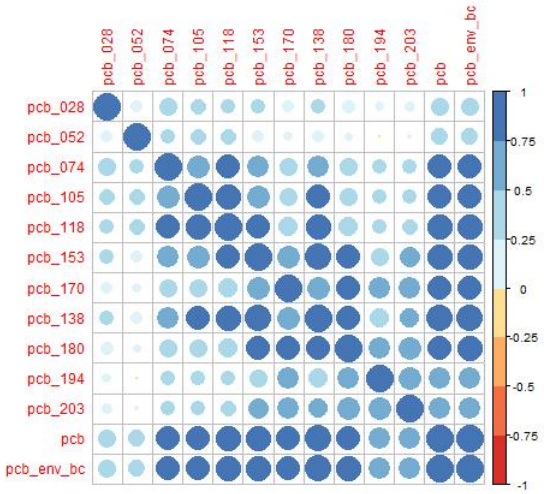


Figure 2: Gestational outcome per hospital center.

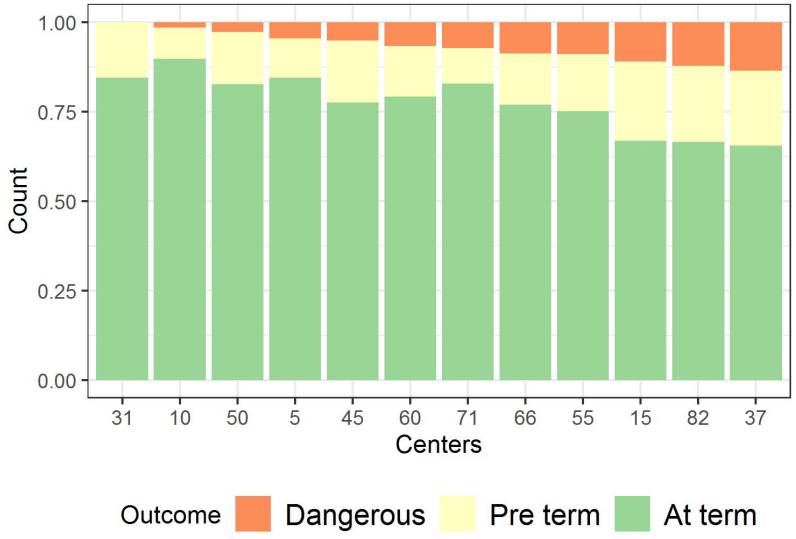


Figure 3: Distribution of estimated exposure to PCB and DDE per gestational outcome and per race.

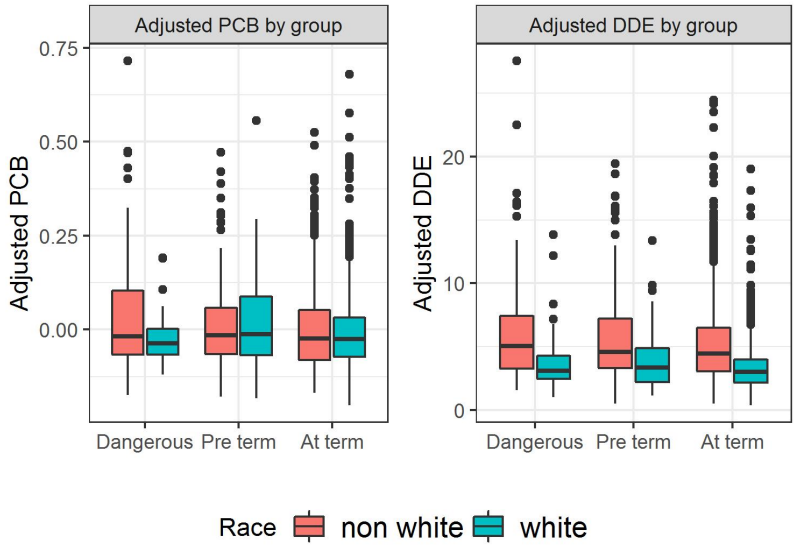
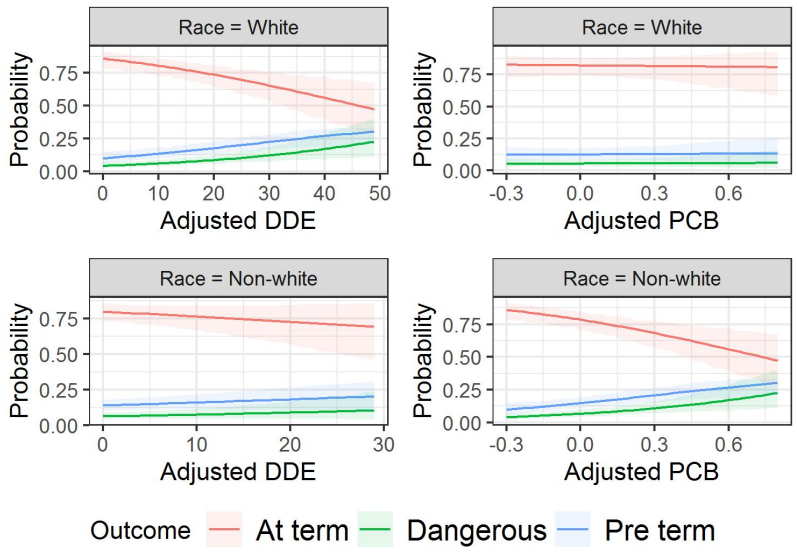


Figure 4: Estimated probability of gestation outcomes in function of race, and exposure to DDE and PCB.



	mean	95%	5%
adjDDE	-0.02	0.01	-0.05
adjPCB	-1.87	-0.81	-2.87
adjDDE*white	-0.06	0.02	-0.12
adjPCB*white	1.69	3.36	0.08

	mean	95%	5%
adjDDE	-0.02	0.01	-0.05
adjPCB	-1.94	-0.88	-2.96
adjDDE*white	-0.06	0.01	-0.13
adjPCB*white	1.77	3.60	0.04

	mean	95%	5%
adjDDE	-0.02	0.01	-0.05
adjPCB	-1.88	-0.82	-2.90
adjDDE*white	-0.06	0.02	-0.13
adjPCB*white	1.74	3.47	0.02

Table 1: 90% credible intervals for all coefficients, under prior on R^2 with location 0.3 (Left), under prior on R^2 with location 0.5 (Middle) under prior on R^2 with location 0.8 (Right)

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

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

Appendix B. Box-Cox analysis for lipid adjustment.

Part of the issue with the exposures of interest in our study (DDE and PCB) is that the substances are lipophilic. This may require to adjust their measurement by the total serum lipid concentration in the blood, so to have an estimate for the excess exposure that comes from the environment. The work by Li et al (2013) suggests a possible correction based on a Box- Cox analysis. In particular, let s_i be the measure for the total lipids serum concentration, and x_i the exposure. The adjusted exposure can be computed by setting

$$x_i^* = x_i / g(s_i) \tag{4}$$

where g is a function to be estimated. A way to do this is by letting g being equal to the Box-Cox correction, that is

$$g(s_i, \lambda) = \begin{cases} \frac{s_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(s_i) & \lambda = 0 \end{cases} \tag{5}$$

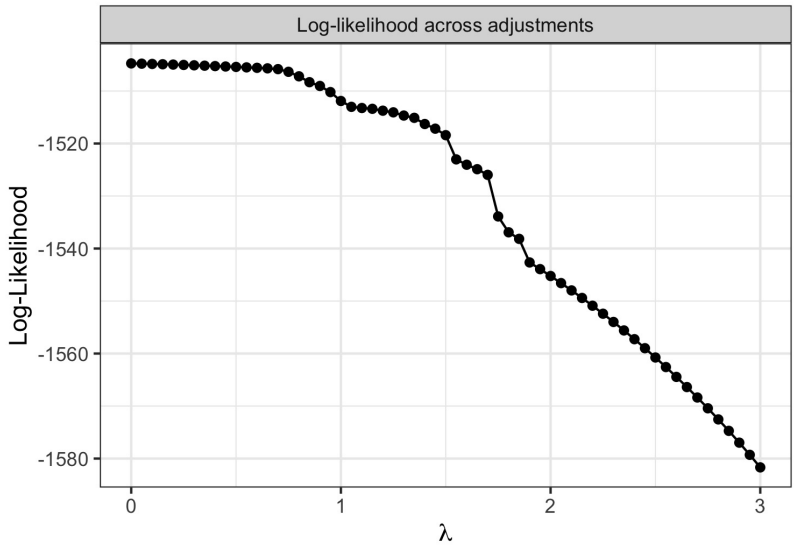
Assuming that there is a unique λ correction for each level of chemical exposure, we can plot the Log-Likelihood across varying levels of λ, and then choose the one that maximizes it. In such a way, we can get rid of the potential case in which serum lipids do not have any impact on the covariate. Under such a scenario, the likelihood should pick at a λ that minimizes the effect of lipids (making the effect of x_i and x_i^*) practically identical.

Following the above reasoning, we plot the Log-Likelihood across varying levels of λ under the transformations

$$\text{DDE}_{\text{exposure}} = \frac{DDE}{g(\text{lipid})} \qquad \text{PCB}_{\text{exposure}} = \frac{PCB}{g(\text{lipid})} \tag{6}$$

We can see that the value at which the log likelihood peaks is 0. This suggests that a log-transformation o both variables in preferable. Note that we do not consider any negative transformation for interpretability reasons.

Figure 5: Log-likelihood for different values of λ.



Appendix C. 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.

Appendix D. Model Checking

Since the ordinal data is used, the common residual model checking plot is no longer applicable. Instead, the surrogate residual method suggested by () is used.

Latent variables Z can be used to parameterize the Bayesian logistics model. Specifically, $Z = -X\beta + \epsilon$ and $Y = j$ if $Z \in [\alpha_{j-1}, \alpha_j]$, where ϵ is a random variable with cumulative distribution $G(\cdot)$ and α_j is some threshold value. $G^{-1}(\cdot)$ is the link function of the model. 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 Z given observation Y. If the model assumptions are 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 empirical distribution of R_S resembles an explicit distribution that is related to the link function $G^{-1}(\cdot)$. Specifically, $R_S \sim G(c + \int udG(u))$ and R_S is independent of X, where c is a constant.

To explain more straightforwardly, if the model assumptions are satisfied, R_S should distribute evenly around 0, independent of X. Besides, the empirical quantiles of R_S should match those of the theoretical distribution.

The scatterplot (Figure6) 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.

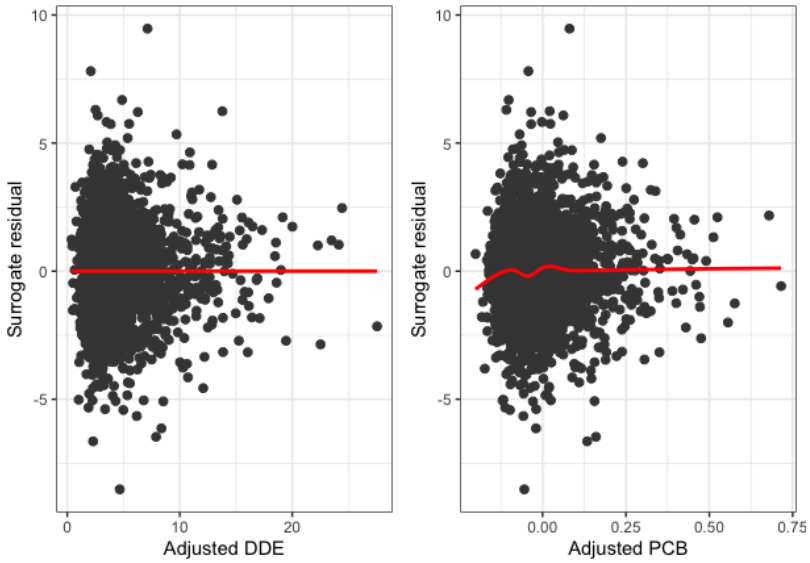


Figure 6: Surrogate residuals of DDE and PCB

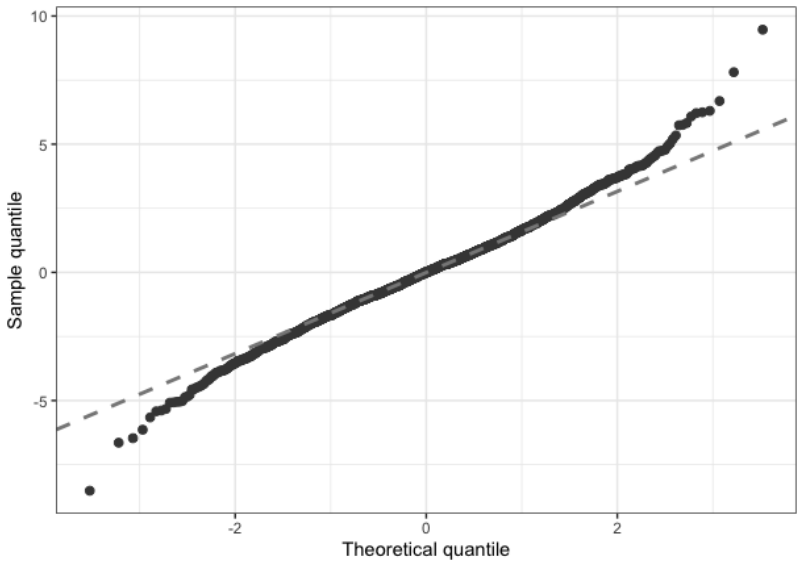


Figure 7: QQ plot of the Surrogate residuals

	5%	95%	mean
dde_env_bc	-0.05	0.01	-0.02
pcb_env_bc	-2.73	-0.75	-1.76
race_aggwhite	0.10	0.85	0.47
center10	-0.13	0.77	0.31
center15	-1.11	-0.28	-0.69
center31	-0.25	0.80	0.26
center37	-1.01	-0.32	-0.65
center45	-0.42	0.35	-0.04
center50	-0.54	0.23	-0.16
center55	-0.78	0.08	-0.35
center60	-0.66	0.16	-0.25
center66	-0.48	0.19	-0.14
center71	-0.46	0.28	-0.09
center82	-1.09	-0.29	-0.69
smoking_status1	-0.31	0.00	-0.16
dde_env_bc:race_aggwhite	-0.12	0.02	-0.05
pcb_env_bc:race_aggwhite	0.01	3.21	1.61

Table 3: 90% credible intervals for all coefficients, under the uniform prior

Appendix E. Full Model Output

The comprehensive output of our model is also included (See Table E for credible intervals and Figure 8 for the histogram). Although the effects of variables other than $\text{DDE}_{\text{exposure}}$ and $\text{PCB}_{\text{exposure}}$ an are not the focus on this report, we can still interpret the coefficients of variables like intercepts and center.

- Intercept: when a subject is non-white, measured at center 5, doesn't smoke, and exposed to 0 level of DDE and PCB, her 90% credible interval for the risk of dangerous preterm is $\frac{1}{1+e^{[-3.24,-2.54]}} * 100\% = [3.77\%, 7.31\%]$. 90% credible interval for the dangerous preterm or preterm is $\frac{1}{1+e^{[-1.88,-1.22]}} * 100\% = [13.24\%, 22.79\%]$.
- Center: There are clear heterogeneity across centers. Center5 is chosen to be the baseline here. Center 15, 37, 82 are significantly different from the baseline because their 90% credible intervals do not cover 0.

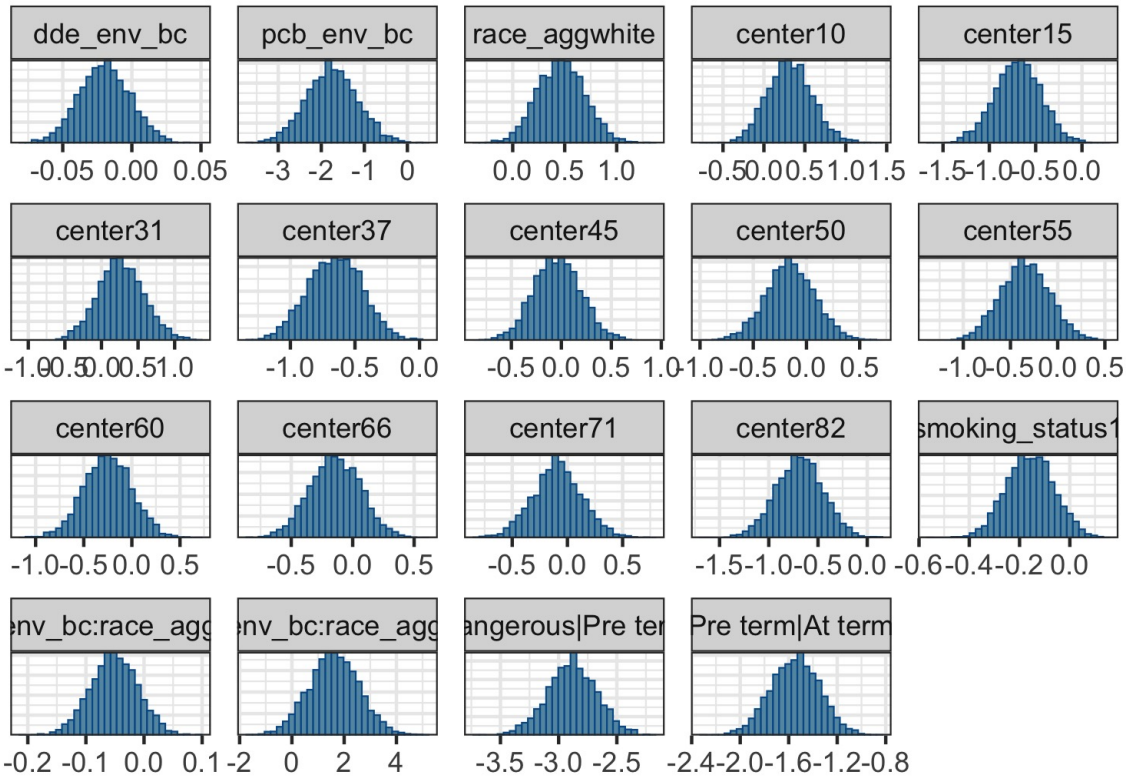


Figure 8: Histogram of 90% credible intervals