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

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

Abstract

This report focuses on the adverse effects of exposures to DDE and PCB chemicals on the chance of premature birth delivery, which may be associated to a weaker health status for the newborn. In particular, we analyze the gestation length of a sample 2380 women of different races and socio-economic status collected across 11 medical centers. We first divide the women into three groups ("Dangerous preterm", "Preterm" and "At term"). As DDE and PCB are lipophilic substances, we correct their concentration measurement to control for an estimate of the lipids in the blood, and later use these adjusted measures as explanatory variables in a Bayesian ordinal logistic regression model with the women's group as outcome. We find that the effects of the chemicals are essentially race dependent: increases in PCB exposure are more detrimental to non-white women, while increases in DDE exposure enhance the odds of a more dangerous delivery on white women.

1. Introduction

Dichlorodiphenyldichloroethylene (DDE) and Polychlorinated Biphenyls (PCB) are two chemical elements that were commonly used in the United States for agricultural purposes and which 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 these elements remain in fatty tissues for a long period of time, studying their impact on human health is particularly important.

In this report, we examine the effect of exposure to DDE and PCB experienced by pregnant women during their lives on birth outcome; more precisely, we assess the potential association between the exposure to these chemicals and the chance of early delivery. We hypothesize that a higher level of exposure to these 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 length of the gestation period. We find that the impact of the substances is essentially race specific: exposure to DDE has a larger impact on the risk of early childbirth among white women while exposure to PCB has a larger impact among non-white women.

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

2. Methods

2.1. Data

The data set consists of $2,380$ pregnant women that visited one of 11 selected american medical centers 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 women's blood, the concentration of cholesterol and triglycerides, the center attended by the woman along with several demographic information (race, level of education, income, occupation, age and smoking status). We remove the 43 women with a length of gestation superior to 45 weeks (which corresponds to the second longest gestation period ever recorded), and the one woman with no value for her PCBs concentrations. Finally, we mean impute the missing data on the income, education and occupational scores[1].

2.2. Feature Engineering

As a first step towards constructing variables relevant to our analysis, we divide the women into three groups, labelled as "Dangerous Preterm", "Preterm" and "At term" based upon the lenght of their gestation (shorter than 33 weeks, between 34 and 36 weeks and longer than 37 weeks, respectively). The groups are meant to capture the danger associated with the birth of the child. 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 level of DDE and PCBs to which the women were exposed during their live. In particular, we calculate the total amount of fatty tissues using the formula suggested in Phillips et al. (1989) and confirmed 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 in order to prevent one measurement from dominating 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 detailed in Section 2.1. We run a forward and backward AIC-based variable selection procedure, 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]. We then fit our Bayesian model on the selected variables with uniform prior via `Stan` using 10 chains of $10,000$ iterations each. The assumptions of the obtained ordinal logistic model are checked in Appendix C.

3. Results

3.1. EDA

Figure 1 presents the correlation matrix of the 11 PCB measurements. We observe that these 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 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 $15, 62, 37$ are pre-term. Figure 3 displays the distribution of the estimated levels of exposure to PCB and DDE per gestational outcome and per race. We can observe a weak negative association between gestation outcome and level of exposure to DDE. We also note that the distribution of the chemicals vary across the two race groups. Figure 2 and Figure 3 indicate that we need to control for race and centers in the regression model in order to prevent these variables from acting as confounders.

3.2. Main Findings

Table 1 reports the mean and 90% confidence intervals for the posterior draws of the coefficients of the DDE and PCB variables and their interactions with the race variable. All the other coefficients, including the center effects, are reported in Table D. Two main facts are worth underlining.

3. This correction derives from a Box-Cox analysis of our model, following the basic procedure in Li et al (2013). See Appendix B 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 wished 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 had to settle for AN AIC-based variable selection 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.

First, we see that the estimates corroborate our hypothesis: an increase in either the PCB or DDE concentration in blood is associated with larger odds of having a dangerous delivery. In particular, a 1 unit increase of $\text{DDE}_{\text{exposure}}$ increments the odds of a more adverse preterm by 2.02% for non whites and by 7.25% for white women[6]. On the other hand, increments by 0.1 units of $\text{PCB}_{\text{exposure}}$ increases the odds of the worst delivery by 19.22% for non-white women, and by 1.595% for white ones. The above percentages highlight the second important result of this analysis, namely, the race dependency of the effect of exposure to the two chemicals. This fact is particularly evident from Figure 4, which displays the variation of the predicted probabilities for the three delivery groups across different levels of PCB and DDE[7]. We see that, as the levels of exposures increase, the probability of delivering at term decreases (with white women more sensitive to DDE and non-white women to PCB).

3.3. Sensitivity Analysis

To further check the robustness of our finding, we test the model with a variety of priors of the model's coefficients. We use a uniform prior and an R^2 prior with three different locations (0.3, 0.5 and 0.8). Table 1 and Table 2 report the estimated means and 90% confidence intervals of the posterior coefficient. We see that the effects do not vary significantly in magnitude across the four priors. For example, a 0.1 unit increase in PCB exposure leads to a 20.56%, 20.68% and 21.41% increase in the odds of the most adverse delivery for non-whites under a 0.3,0.5 and 0.8 location of the R^2 prior, and an increase of 19.28% under the uniform prior. Similar results can be checked for the other coefficients, indicating that our method is not sensitive to the choice of priors.

4. Conclusions and further discussion

This study aims to assess the potential association between exposure to DDE and PCB chemicals and early delivery. The core of our analysis is built upon two building blocks: the estimation of exposure levels from the measured level of lipids (calculated with a linear combination between triglycerides and cholesterol) and the blood concentration of the chemicals, and a bayesian ordinal logistic regression with the delivery time as the outcome (dichotomize into *Dangerous preterm*, *Preterm* and *At term*). We find that the exposures have an adverse effect that is highly dependent on race. These effects remain constant across different choice of priors, indicating that our results are robust. Several strategies could be implemented to improve the quality of the analysis. First, one could do MICE to impute the missing values which, unlike mean imputation, has the advantage of potentially retaining the signal present in variables that have a large proportion of missing values. Moreover, we can extend the full model by modeling the effect of the chemicals in a non-linear way as small levels of exposure are likely to have no effect on human health and we expect the effect to stabilize past a certain threshold. One could also allow for an interaction between the two chemical products. Finally, we can further control for the heterogeneity across centers by modeling the response with a hierarchical structure.

6. Say that β is our coefficient associated to the variable x in the ordinal logistic regression. Then, if the outcomes are ordered from worst to best (like in our example), an increase in 1 unit of x is associated with a variation of the odds of the worst outcome by $(e^{-\beta} - 1) * 100\%$.
7. The curves are computed with reference to center "5" and for non-smoking individuals. Probabilities for DDE variations are predicted holding PCB constant at his mean, and vice versa.

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.

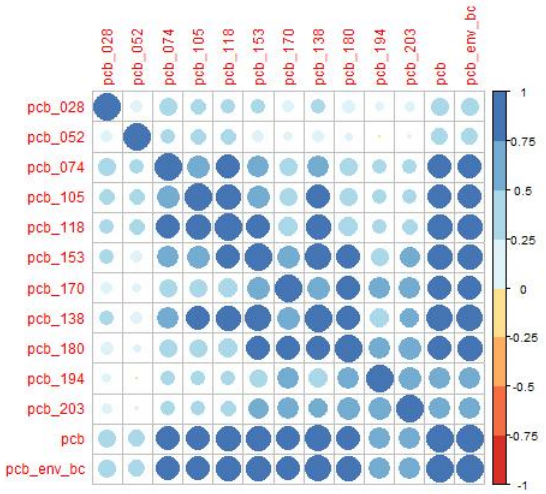


Figure 2: Gestational outcome per hospital center.

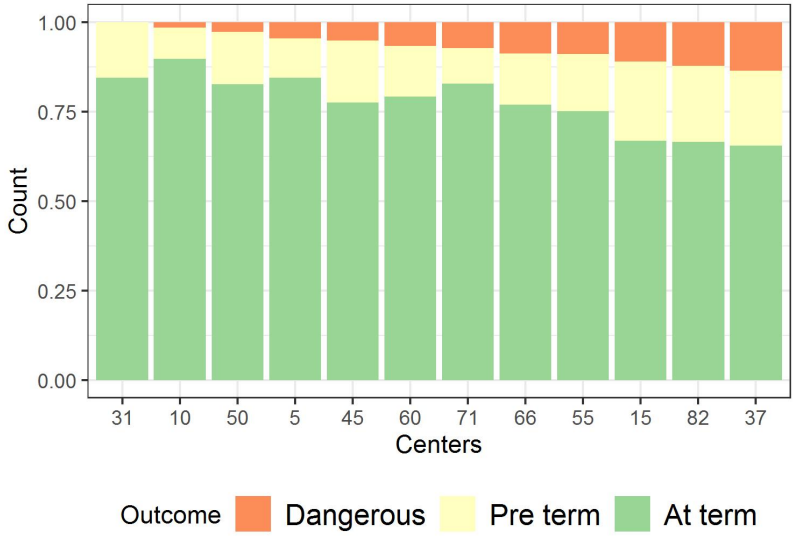


Figure 3: Distribution of estimated exposure to PCB and DDE per gestational outcome and per race.

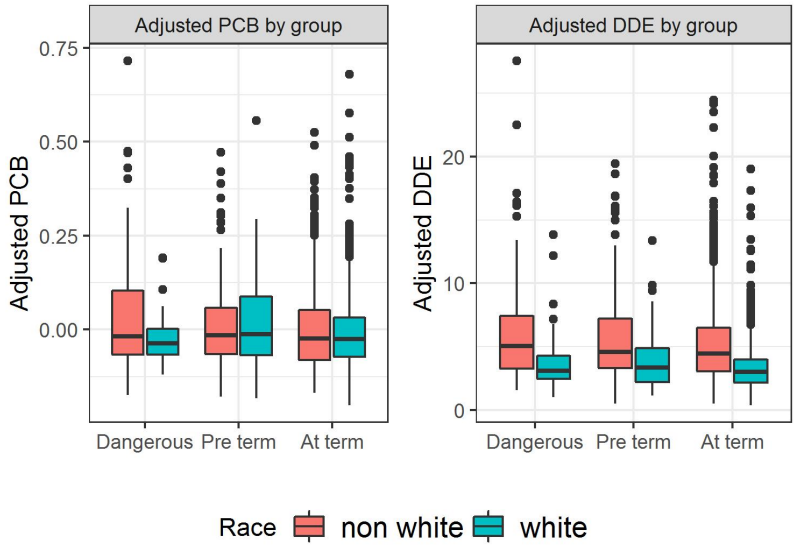
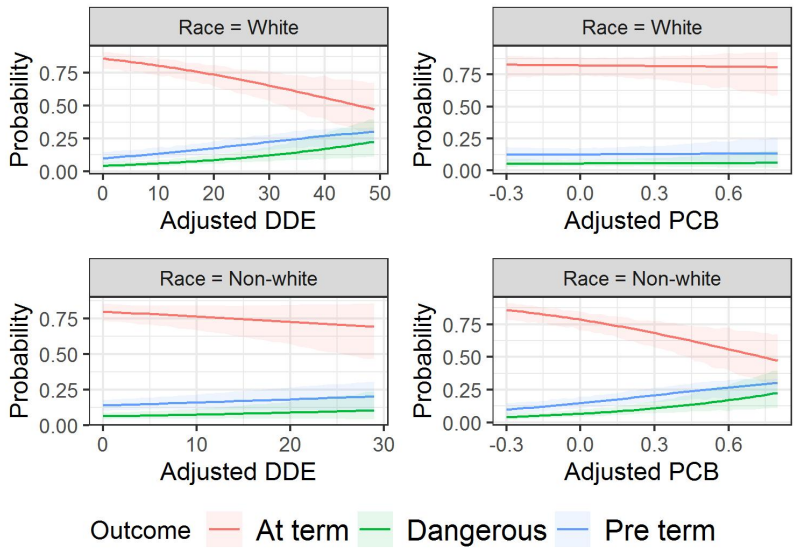


Figure 4: Estimated probability of gestation outcomes in function of race, and exposure to DDE and PCB.



	mean	95%	5%
$DDE_{exposure}$	-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 under a uniform prior.

(a) Location 0.3

	mean	95%	5%
$DDE_{exposure}$	-0.02	0.01	-0.05
$PCB_{exposure}$	-1.87	-0.81	-2.87
$DDE_{exposure}$*white	-0.06	0.02	-0.12
$PCB_{exposure}$*white	1.69	3.36	0.08

(b) Location 0.5

	mean	95%	5%
$DDE_{exposure}$	-0.02	0.01	-0.05
$DDE_{exposure}$	-1.88	-0.82	-2.90
$DDE_{exposure}$*white	-0.06	0.02	-0.13
$DDE_{exposure}$*white	1.74	3.47	0.02

(c) Location 0.8

	mean	95%	5%
$DDE_{exposure}$	-0.02	0.01	-0.05
$DDE_{exposure}$	-1.94	-0.88	-2.96
$DDE_{exposure}$*white	-0.06	0.01	-0.13
$DDE_{exposure}$*white	1.77	3.60	0.04

Table 2: 90% credible intervals and posterior mean of coefficients under three R^2 priors with different locations.

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

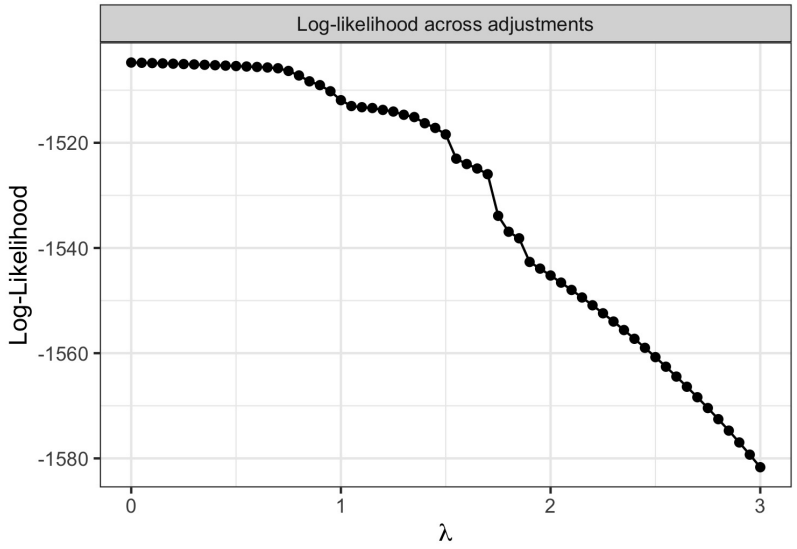
$$g(s_i, \lambda) = \begin{cases} \frac{s_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(s_i) & \lambda = 0 \end{cases} \tag{5}$$

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

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

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



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 since g should be nondecreasing according to our assumptions.

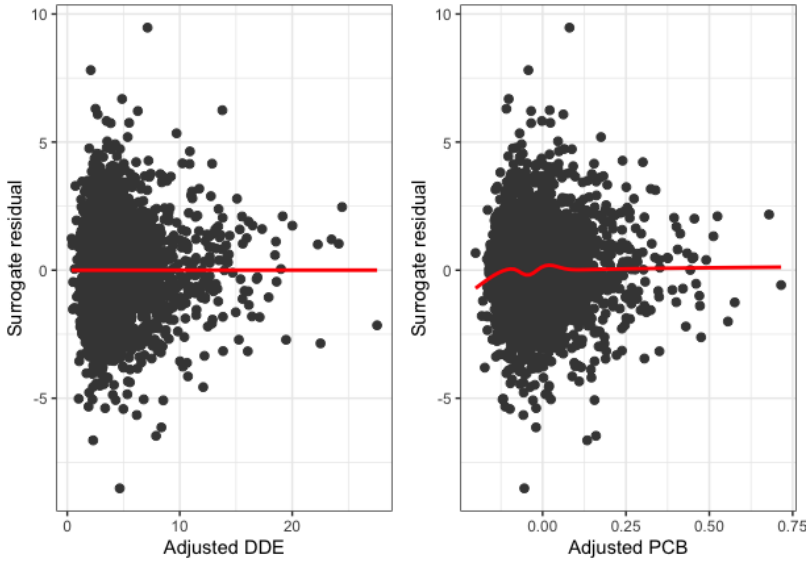


Figure 6: Surrogate residuals of DDE and PCB

Appendix C. 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 Liu and Zhang (2018) 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 (Figure 6) 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.

Appendix D. Full Model Output

The comprehensive output of our model is also included (See Table D 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.

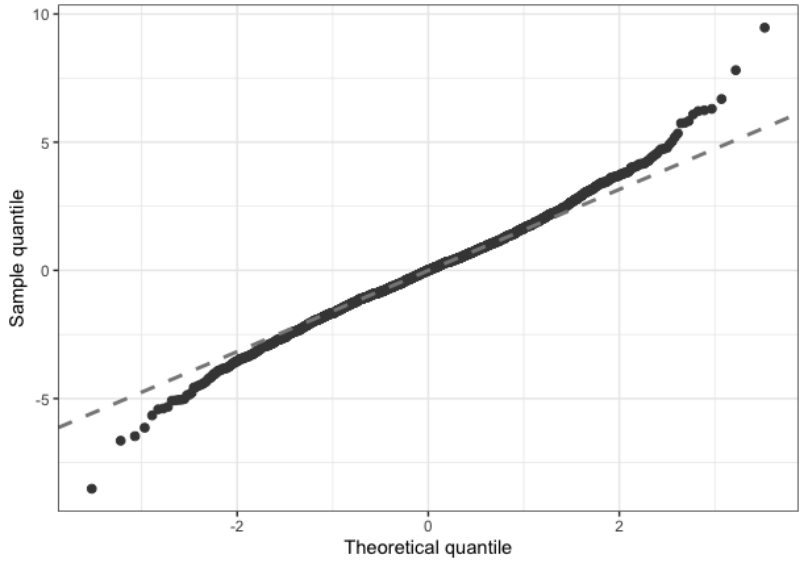


Figure 7: QQ plot of the Surrogate residuals

- 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. Center 5 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.

	5%	95%	mean
$DDE_{exposure}$	-0.05	0.01	-0.02
$PCB_{exposure}$	-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
$DDE_{exposure}$*white	-0.12	0.02	-0.05
$DDE_{exposure}$*white	0.01	3.21	1.61

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

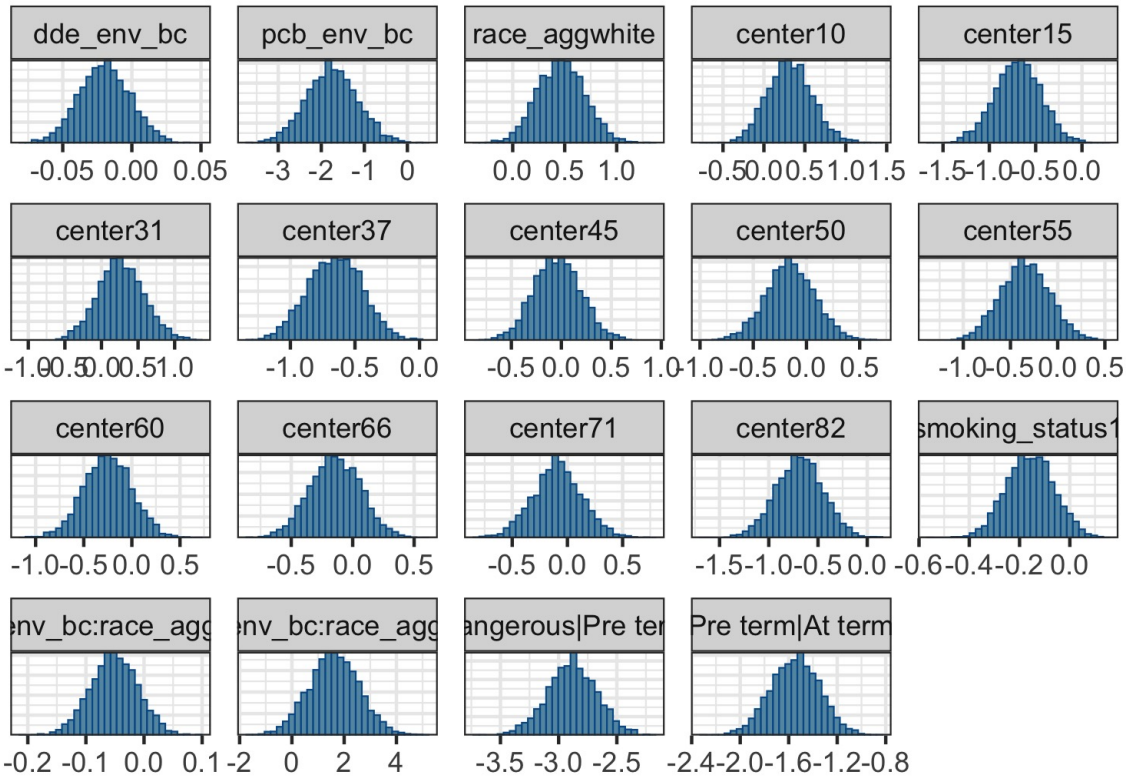


Figure 8: Histogram of posterior draws.