Case study 1: Effect of chemical exposures on preterm birth

Olivier Binette, Brian Kundinger and Joe Mathews

abstract...

1. Introduction

2. Materials and Methods

2.1 Data

Our analysis is based on the *Longnecker* dataset (Longnecker et al., 2001), which contains information on a subset of 2380 pregnant women enrolled in the National Collaborative Perinatal Project (CCP). It relates gestational age to chemical exposures (*dde* and *pcb_** concentrations), to socio-economic variables (*maternal age*, *race*, *smoking status*, *center* of enrollment, and education, income and occupation *score_** variables), as well as to health-related variables (*cholesterol* and *triglycerides* concentrations).

Cleaning up the data, we removed the *albumin* variable, which contained 93% missing values, and dropped observations of pregnancy over 55 week. The single case with missing *pcb_** values was also removed. We are left with $n = 2374$ observations and only missing values in the *score_** variables (22%). *Preterm* birth is defined as gestational age strictly less than 37 weeks ($n = 361$; 15%), and *very preterm* is defined as gestational age strictly less than 34 weeks ($n = 103$; 4%).

To facilitate data visualization and interpretation of the results, we grouped together the "black" ($n = 1220$) and "other" ($n = 123$) race categories into the single "non-white" category. For similar reasons, we summed together the 11 positively correlated *pcb_** variables, obtaining the *totalpcb* variable. While this preserves units, important information may be lost through this process. Other supervised and unsupervised approaches to summarizing the *pcb_** variables are discussed in the appendix.

Challenges and limitations. Our scientific understanding of this data and of the data collection process is limited by a lack of documentation. It not possible to identify the represented population without knowledge of the enrollment mechanism, and therefore our observations may not generalize. Futhermore, it is unclear what should be taken as a scientifically meaningful predictor which strongly correlates to environmental exposures. Should we directly consider *dde* or should it be normalized by lipid concentration? This is not something we can address through the data. Given these limitations and the fact that we have explored many models with the stated goal of identifying an interaction between *dde* and preterm birth risk, the results of our analysis should be taken as highly tentative and exploratory in nature.

Futhermore, there is a noticeable partial correlation between the *dde* and *totalpcb* variables ($\rho = 0.3$), after controlling for the linear effect of the other predictors (excluding *gestational_age*). This could be the shadow of an unobserved confounding variable: unless *dde* and the *pcb_** are breakdown products of the same exposure, this hints at another factor contributing to higher chemical concentrations in the blood. This factor or other toxic chemicals could be causaly related to *gestational_age*, and this prevents us from singling out any potentially causal effect.

2.2 Missing Values Imputation

The 22% missing values in the *score_** variables were imputed under a missing at random assumption using a standard Bayesian framework. We treated the score variables as independent, identically distributed Normal random variables and fitted a Bayesian linear model using the other predictor variables and a non-informative prior. That is, we regressed the observed values of the *score_** variables onto the other predictor variables and treat missing score values as model parameters, estimated through their posterior mean.

This approach has two major limitations. First, separating imputation from model fitting prevents the propagation of imputation uncertainty, although the Bayesian formulation would allow to combine the two

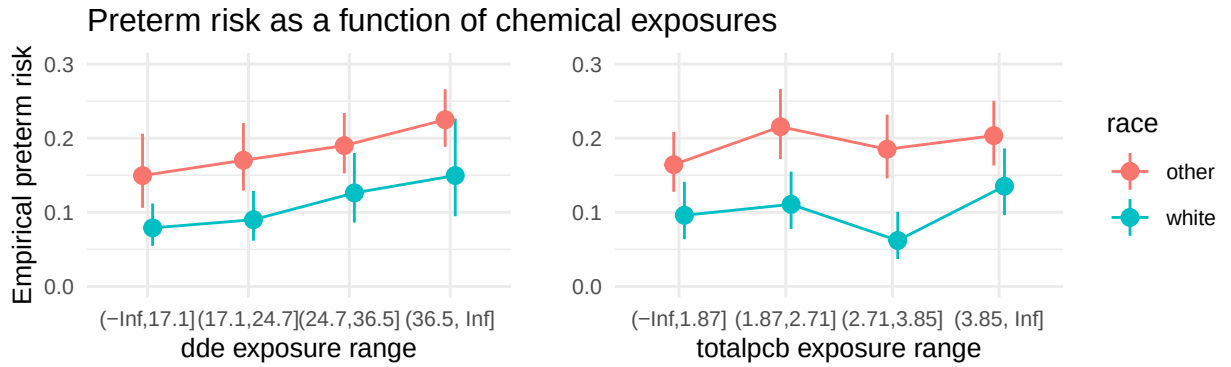

Figure 1: Marginal relationship between empirical preterm risk and chemical exposures, without controlling for socio-economic and health variables. The exposure ranges have been defined so that each class contains 25% of the observations and vertical lines represent 95% confidence intervals for the estimated risk.

steps together. Second, our approach assumes that observed variables in the data set can adequately predict the missing score values, or that data is missing at random. Even if this were to be the case, the data does not suggest clear linear relationships between the *score_** variables and other predictors. The potential of more complex non-linear models for the analysis of this dataset is considered in the Discussion section.

2.3 Logistic regression model for preterm outcome.

We model the log odds ratio of preterm birth as a linear function of the other covariates and perform maximum likelihood estimation. This allows us to control for the baseline effect of the *race*, *smoking_status* and *center* factors, and to control for the linear effect of socio-economic and health-related variables (*maternal_age*, *score_income*, *score_education*, *score_occupation*, *cholesterol*, and *triglycerides*) on the log odds ratio. Appart from *dde* and *totalpcb*, no other variables are incorporated into this base model, and no interaction terms are considered.

The importance of the *dde* and *totalpcb* variables is assessed by: (1) testing for individual and joint statistical significance of the variables; (2) by transforming the associated p-values to the more meaningful $B(p) = -ep\log(p)$ scale; and (3) by interpreting the estimated effect size. For part (1), individual statistical significance is based on standard approximate t-tests and joint significance is tested with an approximate χ^2 test. The rationale for (2) is that $B(p)$ provides a lower bound on the Bayes factor comparing the null hypothesis of no effect to the alternative (over a large nonparametric class of reasonable prior distributions and when $p < e^{-1}$). For example, a p-value of 0.01 is transformed to $B(0.01) \approx 1/8$: if both the null and the alternative hypotheses are a priori equally likely, then a posteriori the alternative hypothesis cannot be more than 8 times more likely than the null. Our interpretation (3) relies on the multiplicative contribution of the *dde* and *pcb* coefficients on the odds ratio of preterm birth, as a function of the empirical quantiles of exposure.

3. Results

Figure 2 below shows the clinically significant estimated effects of the *dde* and *totalpcb* variables. The estimated odds ratio, i.e. the estimated probability of preterm birth to non-preterm birth, can more than double as exposure to *dde* or *totalpcb* goes from zero to some of the larger concentration values observed in the dataset. This is roughly comparable to the marginal effect (which does not take into account the control variables) shown in the left panel of Figure 1. Note that the *dde* and *totalpcb* are correlated ($\hat{\rho} = -0.29$), contributing to the rather large width of the confidence intervals. Furthermore, we warn that the estimated multiplicative effect in the tail may be unreliable. While we expect the logistic regression to properly capture an overall trend, there is not reason for it to adequately fit all features of the data.

Furthermore, the *dde* and *totalpcb* variables have statistically significant effects: it is unlikely that a trend

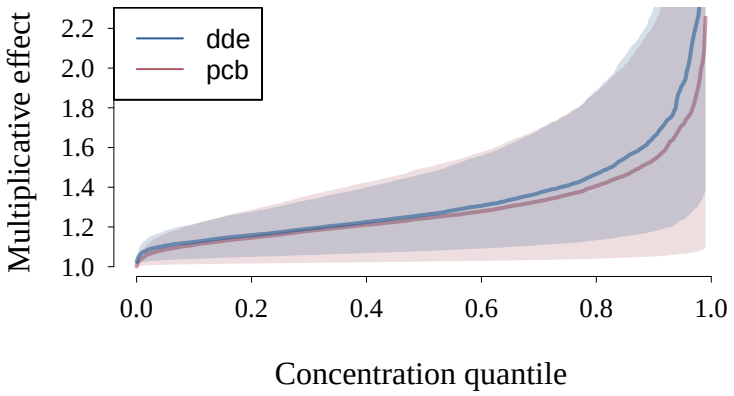

Figure 2: Multiplicative effect of dde and totalpcb on the odds ratio of preterm to non-preterm birth, as a function of the empirical dde or pcb concentration quantile. Confidence intervals are represented by shaded regions, and the plot is truncated on the right at the 0.99 quantile.

of this order would have arised by chance under the logistic model. The p-values computed can be interpreted on the $-ep\log(p)$ scale: the Bayes factor in favor of a non-null effect could be up to 47 to 1 for the *dde* variable, and up to 5.8 to 1 for the *totalpcb* variable.

	dde	*totalpcb*	jointly
p-value	$1.2 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	$2.27 \cdot 10^{-5}$
$B(p)$	$1/47$	$1/5.8$	$1/1514$

These results should be interpreted with care, as we do not expect the logistic model to correctly capture the data-generating mechanism. Inference about the "existence" of an effect is not meaningful in this context, and our analysis should be understood as showcasing a particular aspect of this dataset.

The effect of *dde* and *totalpcb* on preterm risk found in this data, after superficially controlling for confounding factors, should be very concerning. However, we cannot currently single out any of these two variables as having a distinct effect on preterm birth while the partial correlation between them hints at unmeasured confounders.

4. Discussion