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

Raphael Morsomme Rihui Ou Alessandro Zito

Case Study 1 - Stat 723

January 20, 2020

Overview

1. Introduction
2. Data
3. Model (I) - Ordinal Logistic Regression
4. Model (II) - Bayesian Ordinal Logistic Regression
5. Results
6. Conclusions

Introduction

- **Framework**:
  Dichlorodiphenyldichloroethylene (DDE) and *Polychlorinated Biphenyls* (PCBs) are chemicals that persist in the envirnoment and get stored in fatty depositis in the human tissues.
  $\Longrightarrow$ Potential adverse effect on health
- **Question**:
  Is exposure to DDE and PBCs associated with a higher chance of premature delivery in pregnant women?

Pregnancy timeline

- **Dangerous preterm**: delivery at 34 weeks or before (when main organs are underdeveloped)
- **Preterm**: delivery beween 35 and 37 week
- **At term**: delivery after 37 weeks

Data

Data collected by 12 centers contained gestational age (in weeks) of the mother, the DDE and PCBs concentration, socio-economic info and scores (race, occupation, education, income), amount of triglycerides and cholesterol in blood and smoking status.

Preprocessing:

- Drop obs. with gestational age > 45 (the world record)
- Standardize and average levels of PCBs[1]

$$PCB_i = \frac{1}{11}\sum_{j=1}^{11}\frac{PCB_{ij} - mean_i(PCB_{ij})}{sd_i(PCB_{ij})}$$

- Mean impute of occupation, education and income scores
- Aggregate race into $race = 1$ if white and $race = 0$ if non-white

$\Longrightarrow$ Total obs. $=$ **2336**

[1] This avoids the correlation between the PCBs. See the appendix.

Data

- **Our dependent varible is**:

$$gestgroup_i = \begin{cases} 0 & \textit{if} \text{ Dangerous preterm} \\ 1 & \textit{if} \text{ Preterm} \\ 2 & \textit{if} \text{ At term} \end{cases}$$

- To account for triglyceredes and cholesterol, we introduce an **adjusted measure for *PCB* and *DDE*** by:

1. Computing total lipids using Phillips et al.(1989) and Bernert et al.(2007) forumula

$$lipid_i = 2.27 * cholesterol_i + triglycerides_i + 0.623$$

2. Setting[2]

$$adjDDE_i = \frac{DDE_i}{log(lipid_i)} \qquad adjPCB_i = \frac{PCB_i}{log(lipid_i)}$$

[2] The choice of the log comes from a Box-Cox analysis of the log-likelihood, as in Li, Longnecker and Dunson (2013)

EDA (I) - Exposures and gestational groups by race

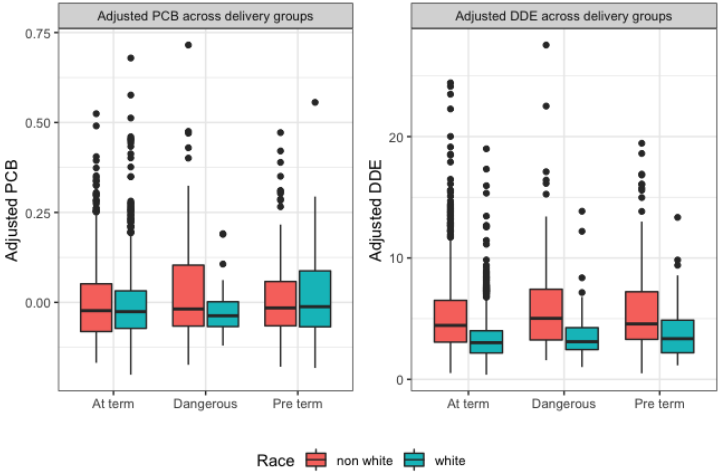

Figure: Relationship between delivery group and adjusted exposures, by race

EDA (II) - Exposure across centers

Model (I) - Ordinal Logistic Regression

After an AIC backward variable selection procedure, our final model is:

$$\begin{aligned} textrmlogit(P(gestgroup_i \leq j)) = \beta_{0j} - \eta_1 adjDDE_i - \eta_2 adjPCB_i \\ - \eta_3 race_i \\ - \eta_4 adjDDE_i * race_i - \eta_4 adjDDE_i * race_i \\ - \sum_{j=center} \eta_{3,j} center_{j,i} + \eta_4 smoke_i \eta_4 adjDDE_i \boldsymbol{\xi}^T \mathbf{z}_i + \varepsilon_i \end{aligned}$$

where

- $j = 0, 1, 2$ is the outcome level
- DDE_i and PCB_i are the amount of DDE and PCB
- $lipid_i$ measures the lipid deposit
- $\mathbf{z}_i$ is a set of covariates.

After an AIC backward , we determine that $\mathbf{z}_i = (center_i, score\_education_i)$
Model assumptions are checked in the appendix.

EDA (II) - Exposure across centers

$$\operatorname{logit}(P(gest_i \leq j)) = \beta_{0j} - \mathbf{X}\beta_i + \varepsilon_i$$

where $j = 0, 1, 2$ corresponds to the outcome level, and $\mathbf{X}$ contains:

- DDE, PCB, race, center, smoke, the 3 scores [main effects]
- (DDE + PCB) * (race + center) [interactions].

AIC-based backward variable selection:

- DDE, PCB, . . . , (PCB + DDE) * race
- (DDE + PCB) * center is not retained

Model assumptions are checked in the appendix.

Model (II) - Bayesian Ordinal Logistic Regression

Results

Conclusions

Appendix (I) - More EDA

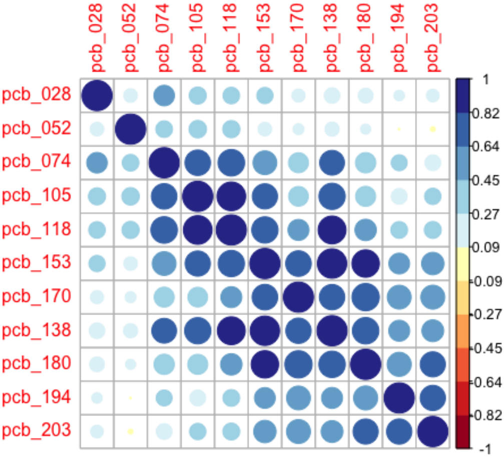

Figure: Correlation plot across PCBs

Model Checking

We can check the assumption of the (frequentist) ordinal logistic model by looking at the Surrogate residuals. If the model assumptions are correct, then the surrogate residuals R_S will have three properties:

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