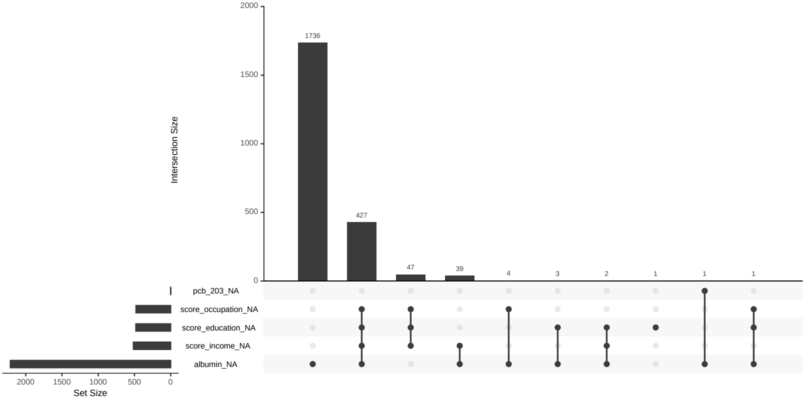
Effect of DDE and PCB Exposure on Pre-Term Delivery

Youngsoo Baek, Yunran Chen, Xiaojun Zheng

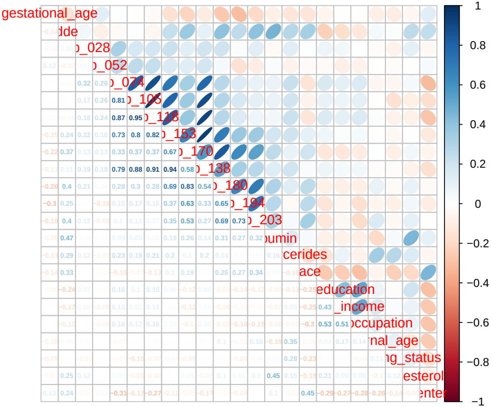
EDA for missing data

- Over 90% missing `albumin`, around 20% missing `score_*`, others less than 0.1%
- Drop the `albumin` and keep the complete cases



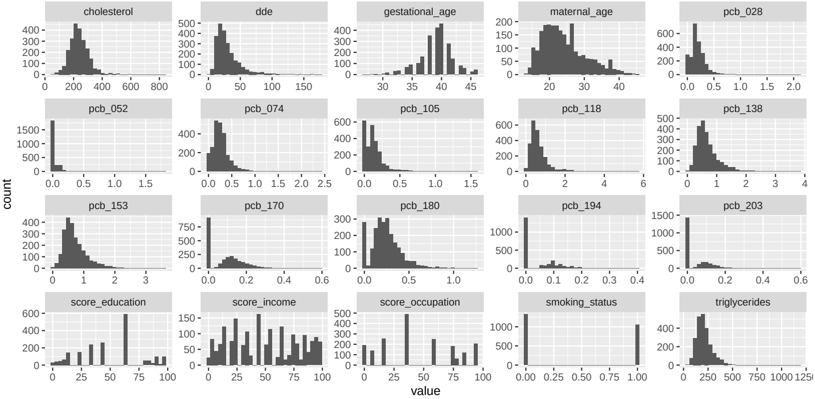
EDA: Correlation Plot

- Weak correlation between covariates and `gestational age`
- Large correlation among covariates especially for PCBs

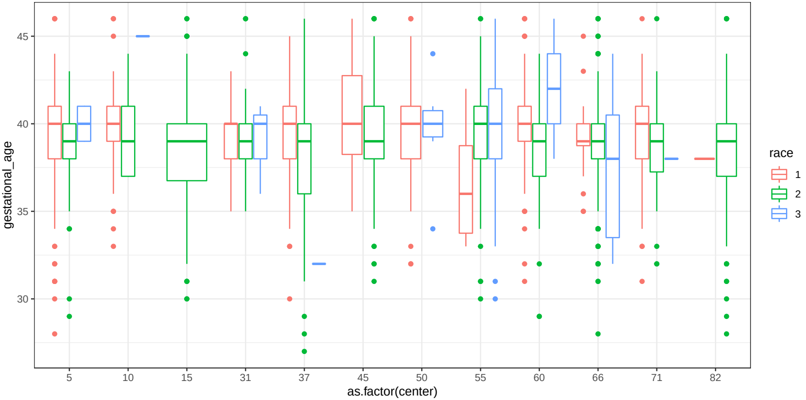


EDA: Histogram for All Variables

- Zero-inflation on some of PCBs
- Long left tail of `gestational age` (after truncated at 46)



EDA: Heterogeneity across Centers



Challenges

- Noisy measurement
- Multicollinearity
- Modeling Heterogeneity between centers

Ultimate goal of the model

What would a hypothetical experimental study for DDE and PCB's look like?

$$\begin{aligned}\text{Null model:}\quad & \text{gestational age} \sim 1 + \text{Demographic variables}\\ \text{Alternative:}\quad & \text{gestational age} \sim 1 + DDE + PCB + ... + \text{Demographic}\end{aligned}$$

Addressing PCB's

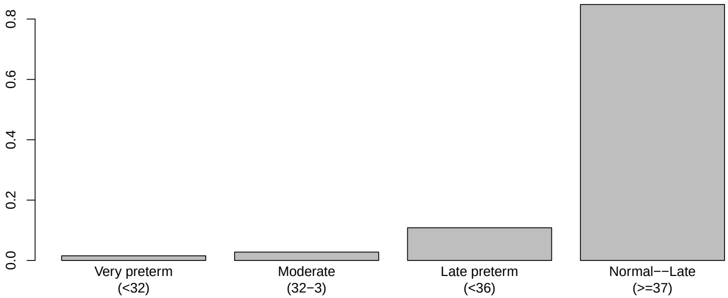
Two main approaches:

i. Everyone contributes, with different weights (**principal component regression**)

ii. Pick a few representative voters (**variable selection**)

Model: Logistic vs. Ordinal logistic

- Binary response: preterm delivery ($<$37 wks)
  - Loss of information about different levels of risk involved in ordered levels



Interpreting the model

- Logistic: coefficients β correspond to $\times e^{\beta}$ increase in the preterm delivery *odds*
- Ordinal logistic: Assumes multiple (>2) delivery category odds are *proportional*

$$\Pr(Y_i \leq k | X_i) = \mathrm{logit}^{-1}(\beta_{0,k} + \boldsymbol{\beta}^T X_i),\ k = 1, 2, 3, 4.$$

- Possible violation: can be proportional, but not by a constant factor

Predictors to be adjusted for

- Excluded: three score variables relating to education, income, and occupation
- Justification: F-test against other predictors excluding chemicals, exploratory model fits
- First principal component for PCB levels (scaled)

Estimated Effects

- 95% confidence interval estimates for significant coefficients

Table 1: Logistic

	Mean	2.5 %	97.5 %
dde	0.009	0.003	0.014
PC1	0.076	0.021	0.130
triglycerides	0.003	0.002	0.005
cholesterol	-0.003	-0.005	-0.001

Table 2: Ordinal logistic

	Mean	2.5 %	97.5 %
dde	0.008	0.014	0.002
PC1	0.081	0.134	0.026
triglycerides	0.003	0.004	0.001

Estimated Effects

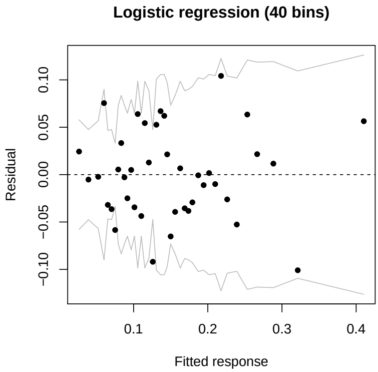
- Two models agree in significant positive mean shifts for center IDs 15, 37, 82 (large number of black subjects)
- "Baseline" log odds +/- 2 standard errors, on probability scale, is estimated for each category by the ordered logistic model.

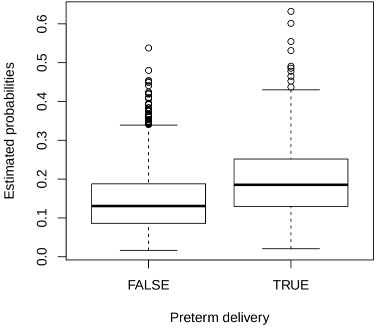
	Lower bound	Mean	Upper bound
Very preterm	0.003	0.008	0.018
Moderately	0.010	0.022	0.049
Late preterm	0.040	0.086	0.174

Interpretation

- Model 1: Adjusted for PCB levels and demographic variables, a 1ug increase in DDE exposure corresponds to 1.009 times more odds of preterm delivery.
- Model 2: (Adjusted) A 1ug increase in DDE exposure corresponds to 1.008 times more odds of more severely preterm delivery (very than moderately so, etc.).
- Similar interpretation can be done for `PC1` and individual weights given to PCB compounds, since the weights are all positive
- However, inference is unidentical to DDE in the sense that we are not adjusting for other PCB compoounds

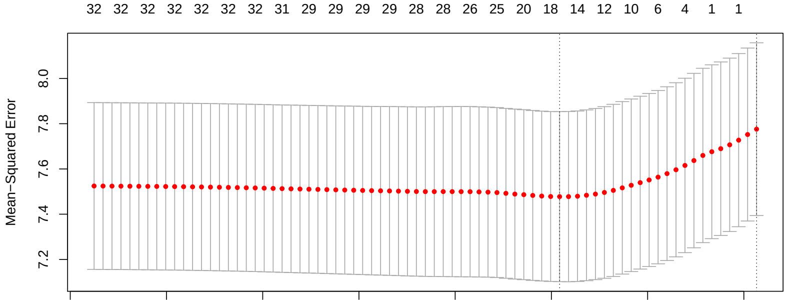
Diagnostic Plots





Further Discussion: PCBs

- Aggregating information of all the PCB's (PCA -> hard to interpretate)
- Selecting representative PCB's (Frequentist and Bayesian variable selection)
    - Bayesian Model Averaging
    - Horseshoe Prior
    - Hierarchical Prior
    - Lasso (Gaussian: dde, pcb_028, 074, 153; Logistic: dde, pcb_074, 153) -> 'consistent' with the previous model



Possible Improvements

- Pooling heterogeneous effects across centers
- Incorporating interactions: systematic, priors-based approach
- Different methods to tackle nonlinearity