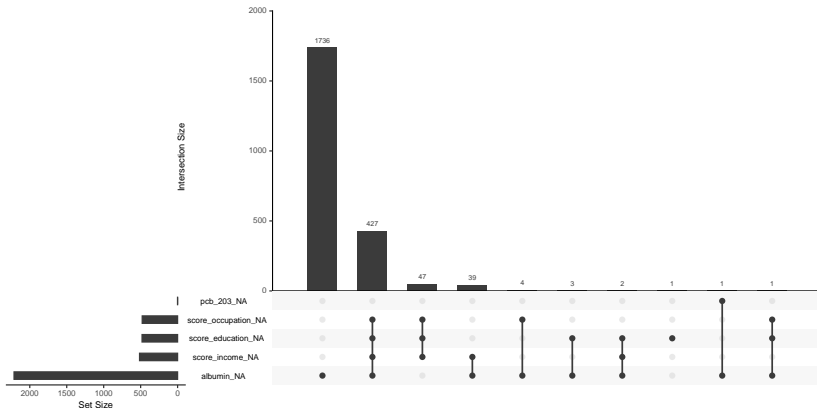


# Effect of DDE and PCB Exposure on Pre-Term Delivery

Youngsoo Baek, Yunran Chen, Xiaojun Zheng

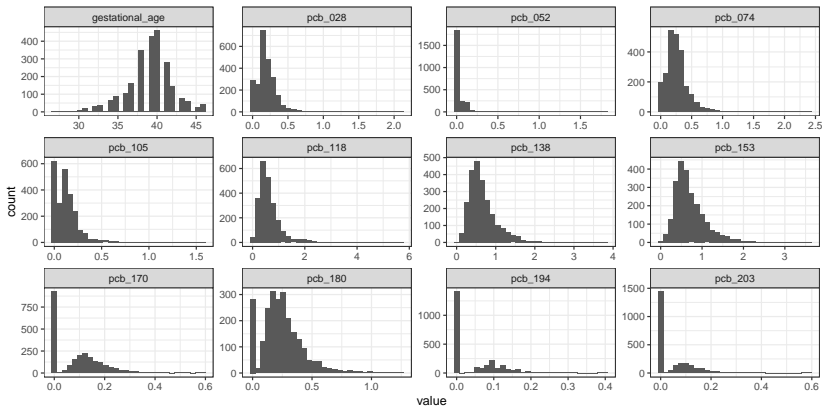
# Missing data

- ▶ Over 90% missing albumin, around 20% missing score\_\*, others less than 0.1%
- ▶ Drop the albumin and keep the complete cases



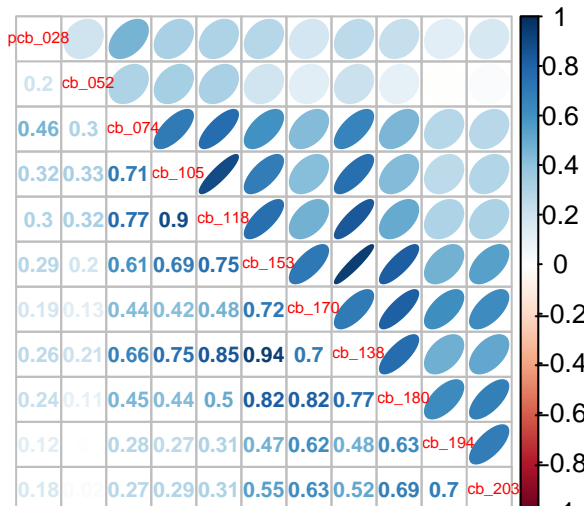
# Noisy measurement

- ▶ Zero-inflation on some of PCBs
- ▶ Long left tail of gestational age + goal -> categorical variable



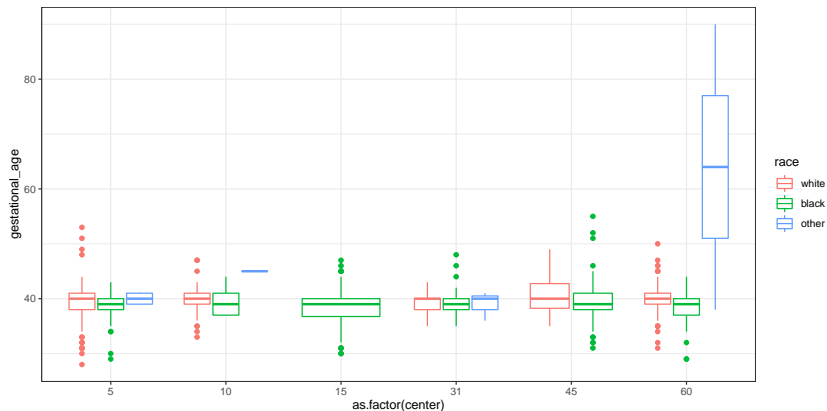
## Multicollinearity

- ▶ Weak correlation between covariates and gestational age (less than 0.3)
- ▶ Large correlation among covariates especially for PCBs -> PCA vs Variable Selection



# Heterogeneity across Centers: fix effects

The demographic background (race, income, etc.) vary across center. -> fix effects vs random effect



## Ultimate goal of the model

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

Null model: gestational age  $\sim 1 + \text{Demographic variables}$

Alternative: gestational age  $\sim 1 + DDE + PCB + \dots + \text{Demographic}$

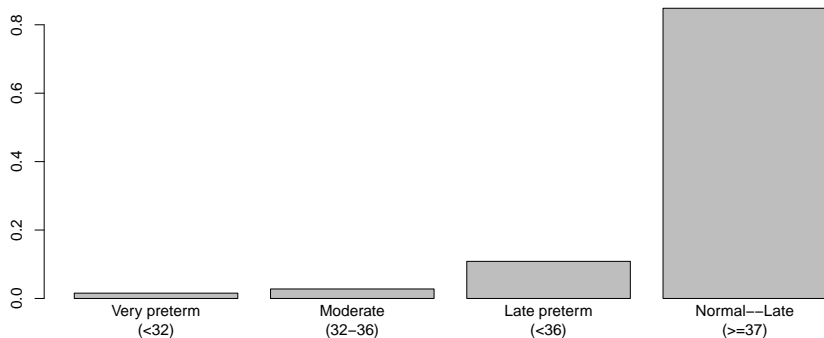
# Addressing PCB's

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

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

- ▶ 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	0.002	0.014
PC1	0.081	0.026	0.134
triglycerides	0.003	0.001	0.004

## Estimated Effects

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

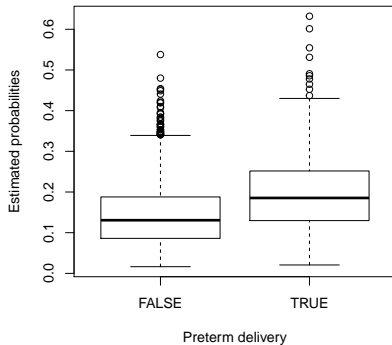
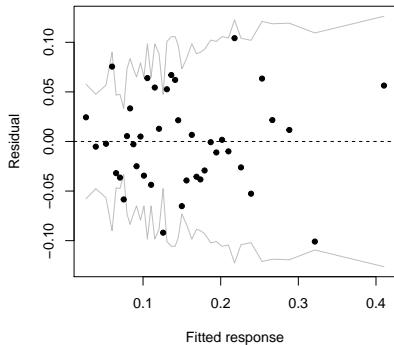
	Lower bound	Mean	Upper bound
Very preterm	0.003	0.008	0.018
Moderately	0.010	0.022	0.049
Late preterm	0.040	0.086	0.174

# 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 compounds

# Diagnostic Plots

Logistic regression (40 bins)



## Further Discussion: Select important PCBs

- ▶ Aggregating information of all the PCBs (PCA -> hard to interpret)
- ▶ Selecting representative PCB's (Frequentist and Bayesian LASSO)
  - ▶ Lasso (DDE, PCB 074, 153 + demographic covariates)

## Estimated Effects (Variable Selection)

Table 3: Logistic

	Mean	2.5 %	97.5 %
dde	0.008	0.001	0.015
pcb_074	0.426	-0.350	1.160
pcb_153	0.368	-0.081	0.811

Table 4: Ordinal logistic

	Mean	2.5 %	97.5 %
dde	0.007	0.001	0.014
pcb_074	0.585	-0.164	1.285
pcb_153	0.340	-0.098	0.771



## Testing for PCBs (Variable Selection)

Table 5: Testing for PCBs in logisitic and Ordinal logistic

Deviance	Df	Pr(>Chi)
8.595	2	0.014
10.429	2	0.005

## Possible Improvements

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