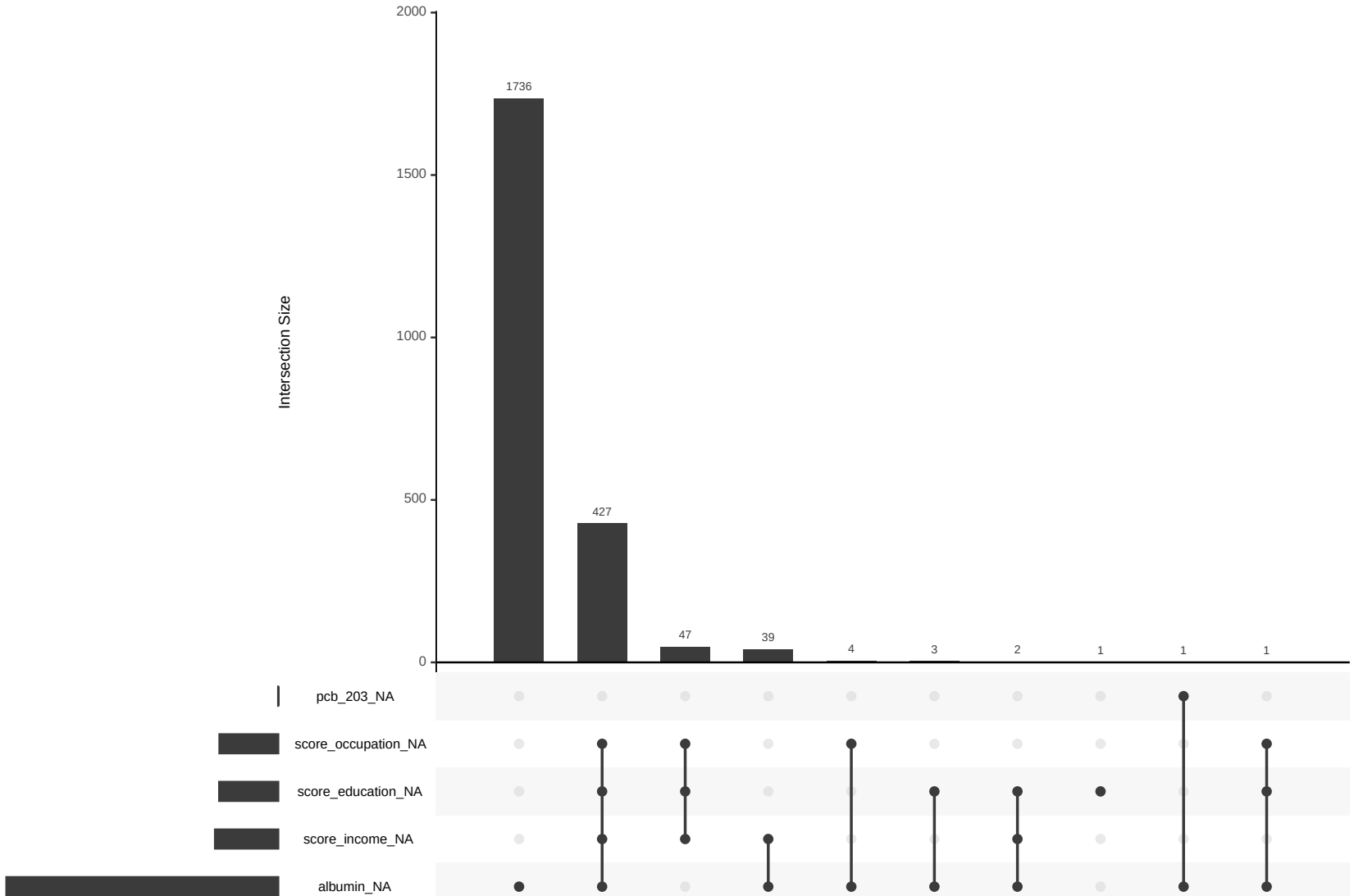
Effect of DDE and PCB Exposure on Pre-Term Delivery

Youngsoo Baek, Yunran Chen, Xiaojun Zheng

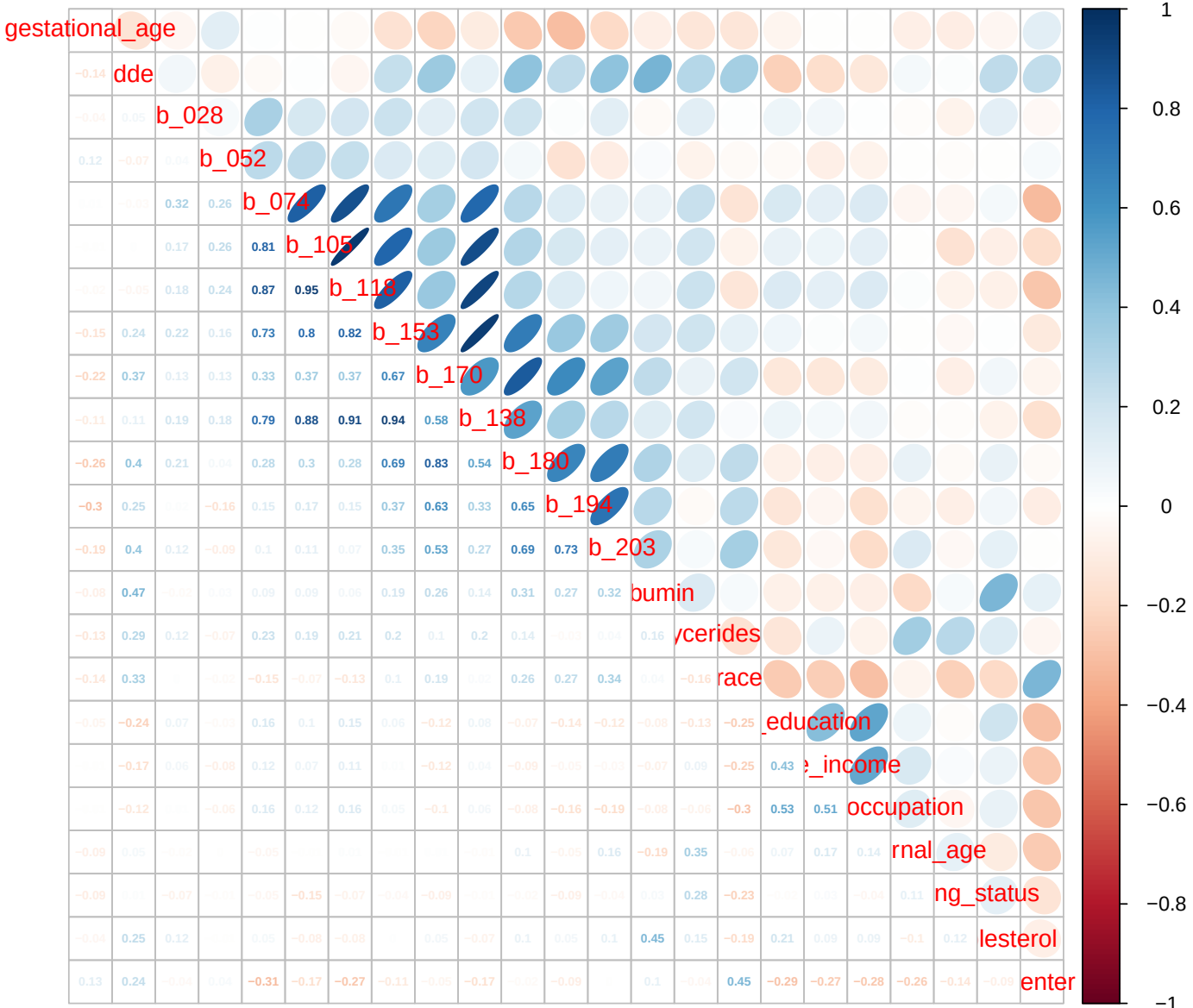
EDA for missing data

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



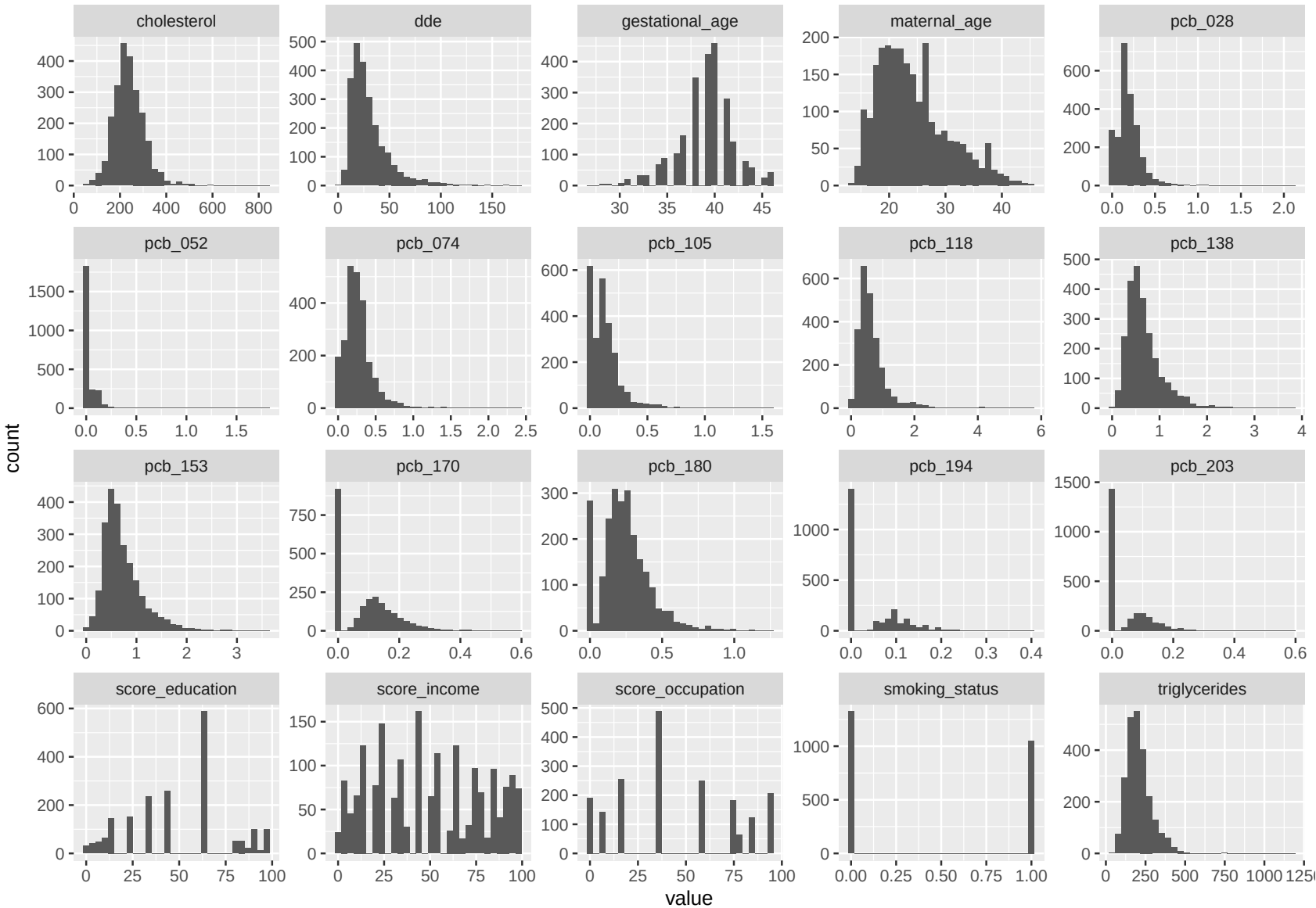
EDA: Correlation Plot

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



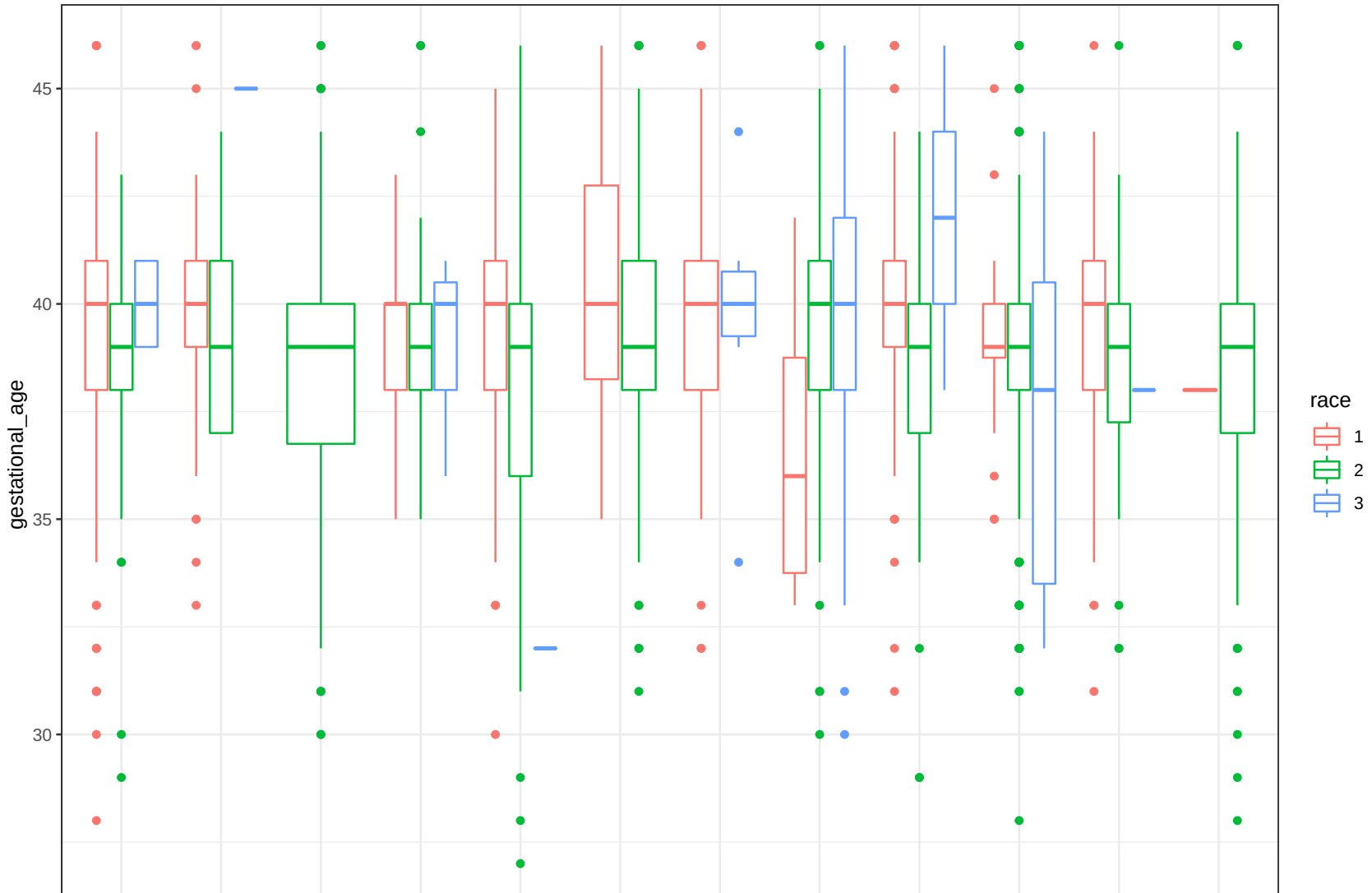
EDA: Histogram for All Variables

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



EDA: Heterogeneity across Centers

The demographic background of each center varies a lot, suggesting heterogeity across centers. (favor fixed effect instead of random effect)



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?

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

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

Model: Logistic vs. Ordinal logistic

- Binary response: preterm delivery (<37 wks)
  - Loss of information about different levels of risk involved in different periods
- Ordinal response: categories of delivery periods
  i. Very preterm ($<$ 32 wks)
  ii. Moderately preterm (32 or 33 wks)
  iii. Late preterm (majority; $<$ 36 wks)

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*

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

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

Predictors to be adjusted for

- Indication from model fits against including score variables
  - Advantage of resolving missingness issues
- First principal component for PCB levels (scaled)
  - Multicollinearity abetted at cost of direct interpretation of effect size

(. . . Want some kind of plot here illustrating the above points . . .)

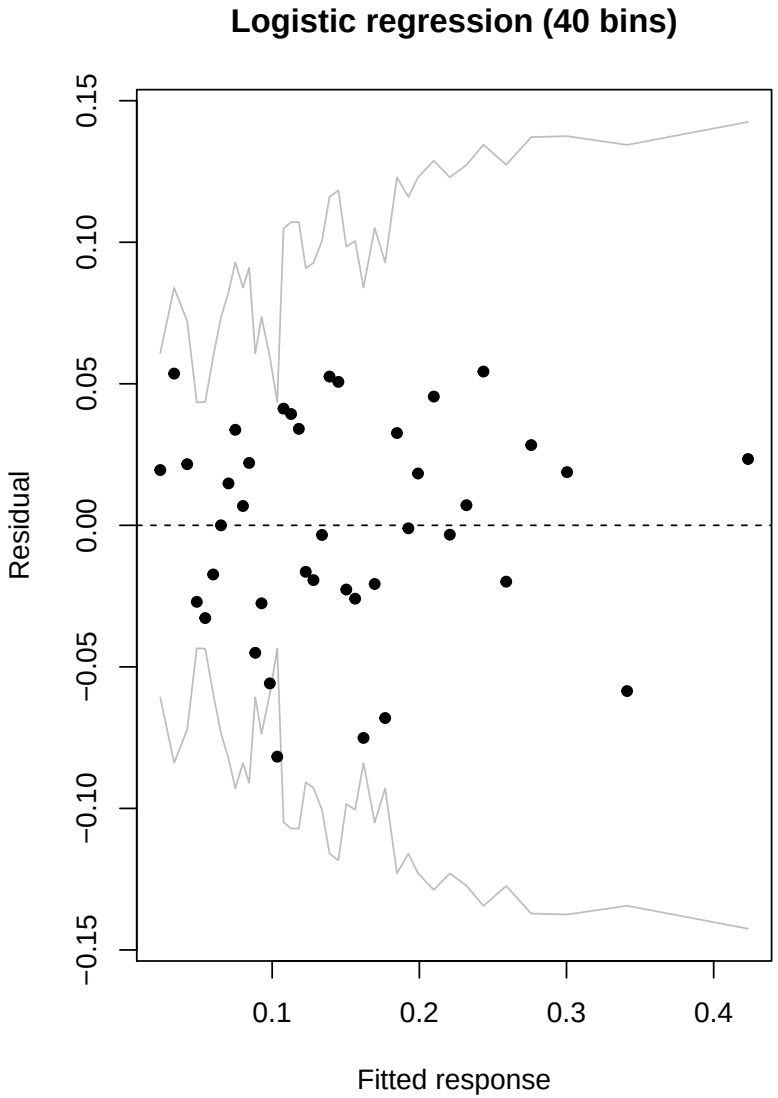
95% (approximate) confidence intervals for the models:

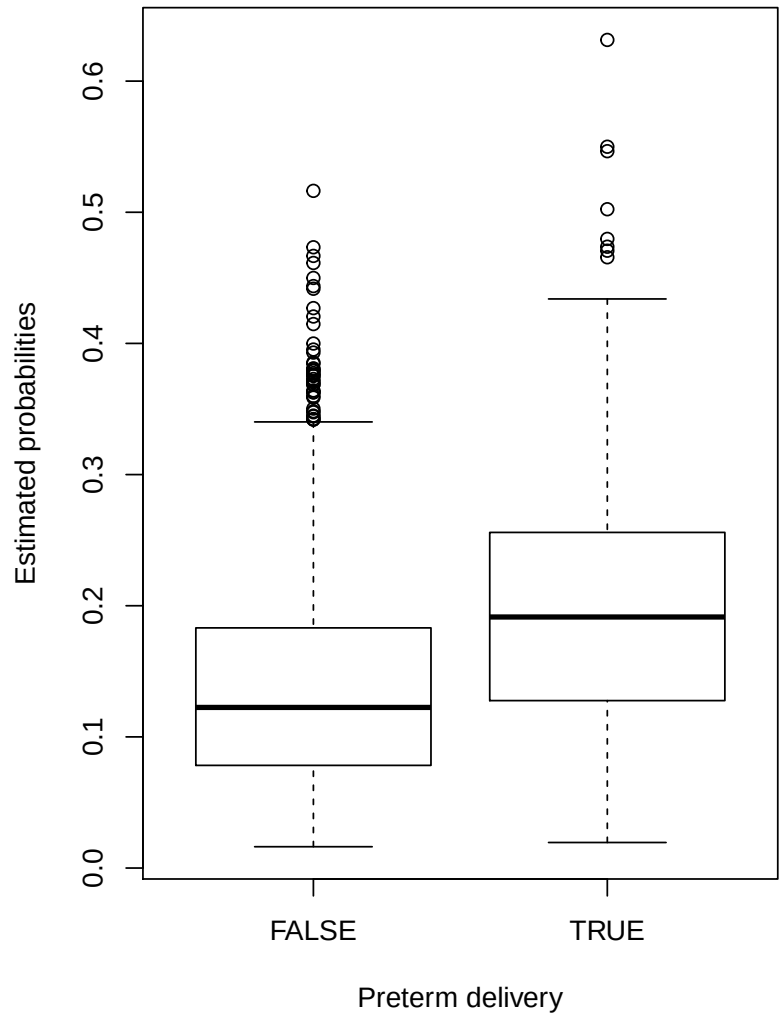
```
##       Mean 2.5 % 97.5 %
## dde 0.009 0.002  0.015
## PC1 0.077 0.014  0.138
```

```
##       Mean 2.5 % 97.5 %
## dde 0.009 0.002  0.015
## PC1 0.077 0.014  0.138
```

- Hard to see improvement from a "baseline" model (Gaussian)

Diagnostic Plots






```
## Analysis of Deviance Table
##
```

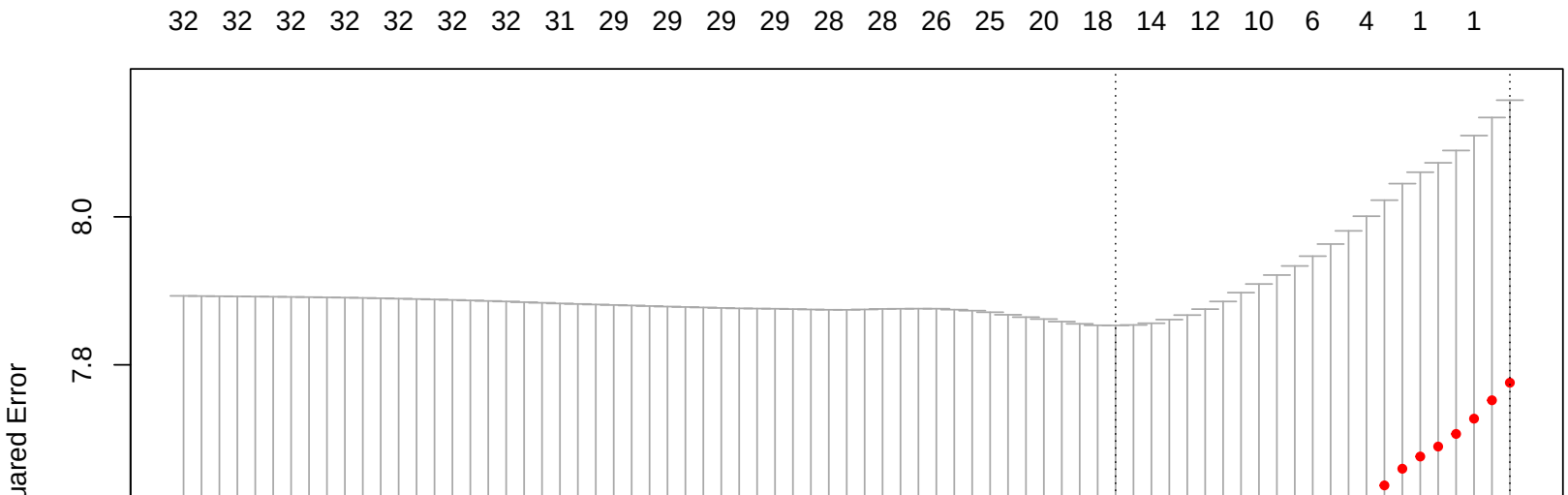
Inference on Effect Sizes

. . . (clear interpretation of the logistic/ordinal coefficients needed here)

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

```
## Loaded glmnet 3.0-1
```



Possible Improvements

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