Case Study 1-Group 1

Melody Jiang, Irene Ji, Keru Wu

Department of Statistical Science, Duke University

01/21/2019

Introduction

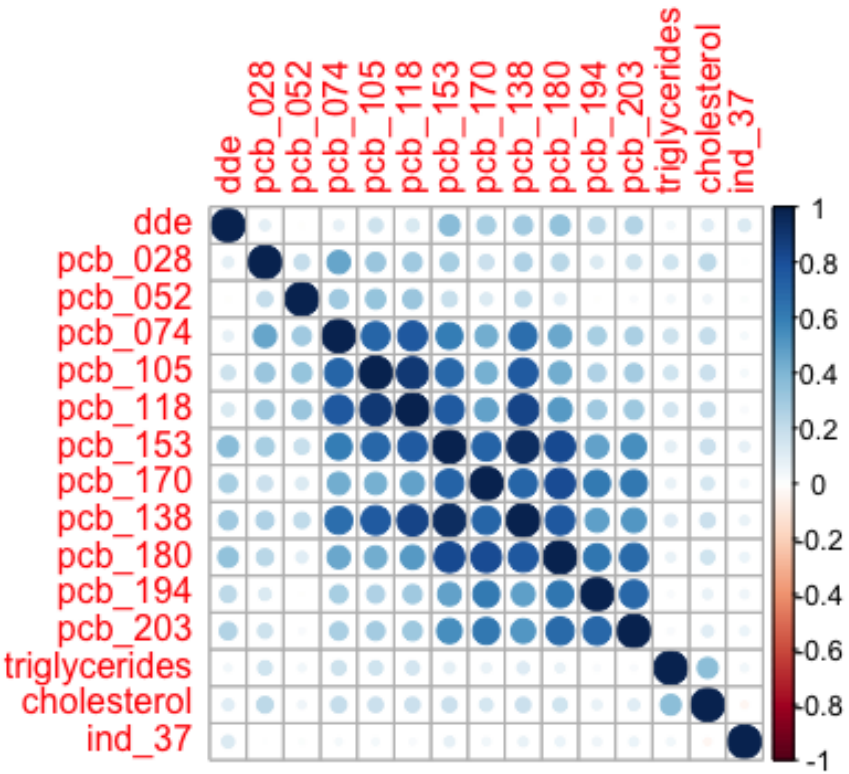
- Data: Subset of National Collaborative Perinatal Project (CPP), comprised of 2380 observations of pregnant women [Longnecker et al., 2001].
- Goal: Assess how DDE and PCBs associate with risk of premature delivery, adjusting for confounding variables.

EDA and Preprocessing

- Premature delivery: Gestational Age ≤ 36.
- Standardize continuous variables.
- Missing data: Multivariate Imputations by Chained Equations (MICE package in R) for covariates. Deleted albumin because 93 percent missing. Only one observation missing in dde and pcb, deleted.
- Limit of Detection (LOD): Exists in some PCBs. All LODs are negligible compared to data scale (e.g. 0.01 compared to 0.3)

EDA and Preprocessing: Collinearity and Dimensionality Reduction

- There are 11 types of PCBs, some of which have high correlation and might distort modeling result.



- Possible approaches: Simple sum, PCA, Factor Analysis.

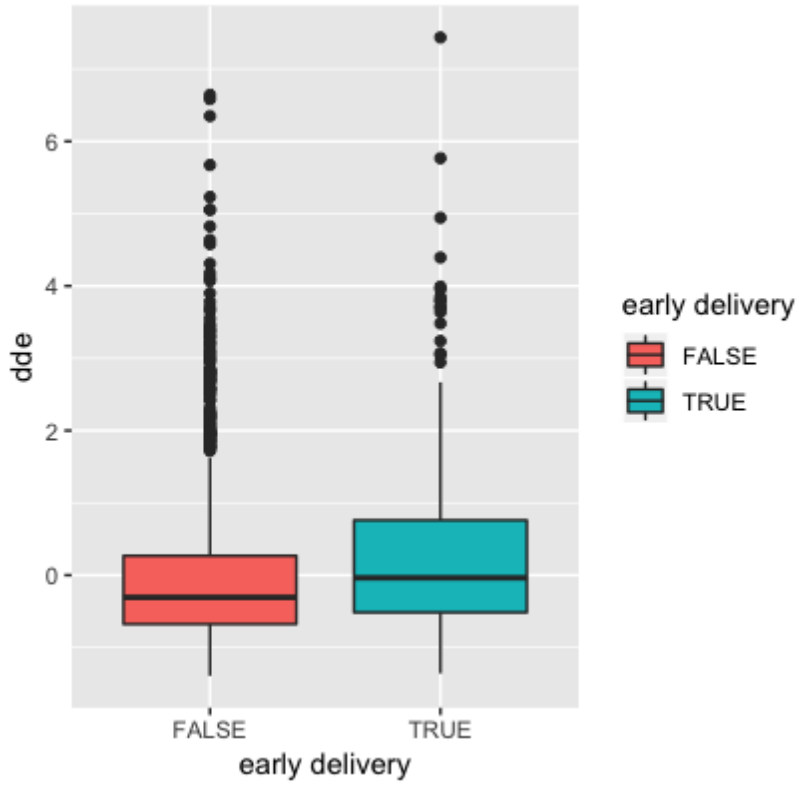
EDA and Preprocessing: Collinearity and Dimensionality Reduction

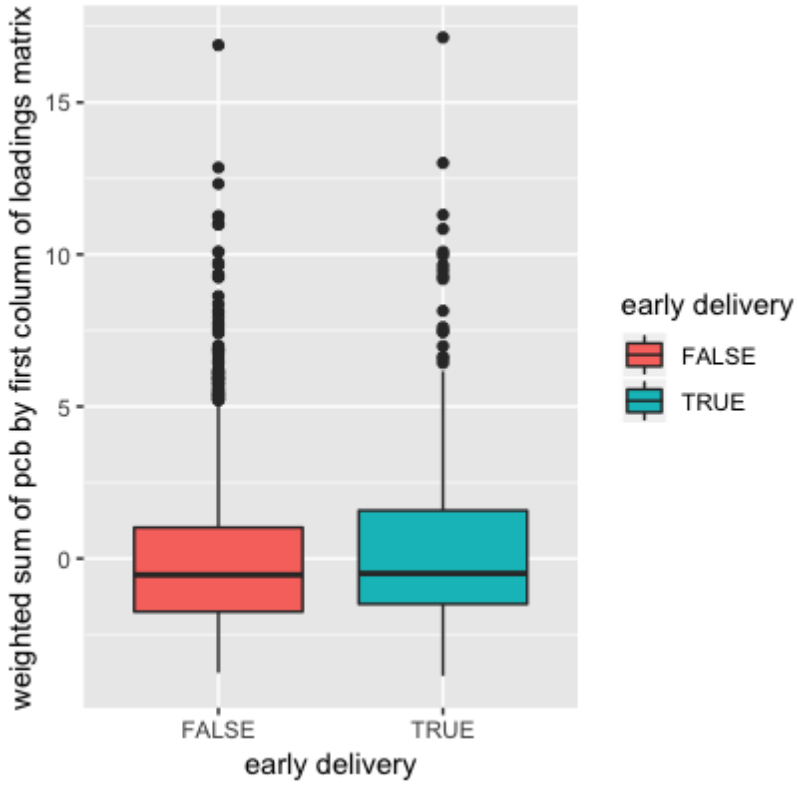
- Possible approaches: Simple sum, PCA, Factor Analysis.

```
Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
pcb_028  0.161  0.243  0.833  0.342  0.154
pcb_052  0.116  0.376  0.223 -0.886
pcb_074  0.306  0.314         0.189 -0.217
pcb_105  0.320  0.333 -0.208        -0.282
pcb_118  0.342  0.306 -0.248        -0.199
pcb_153  0.376         -0.160         0.332
pcb_170  0.325 -0.274         -0.123  0.323
pcb_138  0.383         -0.225         0.165
pcb_180  0.344 -0.277                 0.375
pcb_194  0.253 -0.419  0.158 -0.100 -0.585
pcb_203  0.268 -0.409  0.203 -0.106 -0.290

                       Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.4458646 1.3261098 0.94105657 0.89065865 0.70742028
Proportion of Variance 0.5440699 0.1599370 0.08054181 0.07214604 0.04551399
Cumulative Proportion  0.5440699 0.7040069 0.78454872 0.85669476 0.90220875
```

EDA and Preprocessing





Model

- Linear regression or logistic regression?

Model

- Linear regression or logistic regression?
- Too simple & Can't fit nonlinear trend.

Model

- Linear regression or logistic regression?
- Too simple & Can't fit nonlinear trend.
- Domain knowledge?

Model

- Linear regression or logistic regression?
- Too simple & Can't fit nonlinear trend.
- Domain knowledge?
- Chemicals have no effect when concentration is lower than a bound.

Model

- Linear regression or logistic regression?
- Too simple & Can't fit nonlinear trend.
- Domain knowledge?
- Chemicals have no effect when concentration is lower than a bound.
- Chemicals have constant effect when concentration is higher than a bound.

Model

- Linear regression or logistic regression?
- Too simple & Can't fit nonlinear trend.
- Domain knowledge?
- Chemicals have no effect when concentration is lower than a bound.
- Chemicals have constant effect when concentration is higher than a bound.
- Nonlinear Model

Model

- Generalized Additive Model (GAM)

$$g(Y_i) = \beta_0 + \sum_{j=1}^{m} f_i(x_{ij}) + \sum_{k=1}^{l} \beta_k z_{ik}$$

- Choice of g: probit or logit.
- $x_{.j}$s include DDE, PCBs, maternal age, etc.
- $z_{.k}$s include categorical variables and some confounding variables.

Model

- Frequentist approach may overestimate uncertainty.
- Frequentist GAM may produce a non-significant p-value.

Model

- Frequentist approach may overestimate uncertainty.
- Frequentist GAM may produce a non-significant p-value.
- Bayesian Generalized Additive Model

$$g(Y_i) = \beta_0 + \sum_{j=1}^{m} f_j(x_{ij}) + \sum_{k=1}^{l} \beta_k z_{ik}$$

- Adds priors on the common regression coefficients, priors on the standard deviations of the smooth terms.

Discussion

- Deal with different centers

Discussion

- Deal with different centers
- Approach 1: Bayesian Hierarchical Model

Discussion

- Deal with different centers
- Approach 1: Bayesian Hierarchical Model
- Approach 2: Mixed Effect / Random Effect Model

Discussion

- Deal with different centers
- Approach 1: Bayesian Hierarchical Model
- Approach 2: Mixed Effect / Random Effect Model
- Generalized Additive Mixed Model (GAMM)
- Bayesian GAMM

Discussion

- Specialized prior may give narrower credible intervals.

Discussion

- Specialized prior may give narrower credible intervals.
- Including Interactions: Bayesian Factor Analysis (Ferrari, F. and Dunson, D.B. 2019)