

Case Study 1-Group 1

Melody Jiang, Irene Ji, Keru Wu

Department of Statistical Science, Duke University

01/21/2019

Introduction

- ▶ Data: A study by Longnecker et al. (2001), comprised of 2380 observations of pregnant women.
- ▶ Goal: Assess how DDE and PCBs relate to risk of premature delivery.

Data Preprocessing

- ▶ Preterm birth: Gestational Age ≤ 36 .

Data Preprocessing

- ▶ Preterm birth: Gestational Age ≤ 36 .
- ▶ Standardize continuous variables.

Data Preprocessing

- ▶ Preterm birth: Gestational Age ≤ 36 .
- ▶ Standardize continuous variables.
- ▶ Missing data: Multivariate Imputations by Chained Equations (MICE package in R).

Data Preprocessing

- ▶ Preterm birth: Gestational Age ≤ 36 .
- ▶ Standardize continuous variables.
- ▶ Missing data: Multivariate Imputations by Chained Equations (MICE package in R).
- ▶ Limit of Detection (LOD): Exists in some PCBs. All LODs are negligible compared to data scale (e.g. 0.01 compared to 0.3)

Collinearity & Dimensionality Reduction

- ▶ There are 11 types of PCBs, some of which have high correlation and may be redundant.

Collinearity & Dimensionality Reduction

- ▶ There are 11 types of PCBs, some of which have high correlation and may be redundant.
- ▶ Possible approaches: Simple sum, PCA, Factor Analysis.

Collinearity & Dimensionality Reduction

- ▶ There are 11 types of PCBs, some of which have high correlation and may be redundant.
- ▶ Possible approaches: Simple sum, PCA, Factor Analysis.
- ▶ Alternative approach: model with regularization (e.g. LASSO).

Collinearity & Dimensionality Reduction

- ▶ There are 11 types of PCBs, some of which have high correlation and may be redundant.
- ▶ Possible approaches: Simple sum, PCA, Factor Analysis.
- ▶ Alternative approach: model with regularization (e.g. LASSO).
- ▶ Check Multicollinearity: Variation Inflation Factor (VIF).

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)

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

- ▶ 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)