Case Study 1-Group 1

Melody Jiang, Irene Ji, Keru Wu

Department of Statistical Science, Duke University

01/21/2019

Introduction

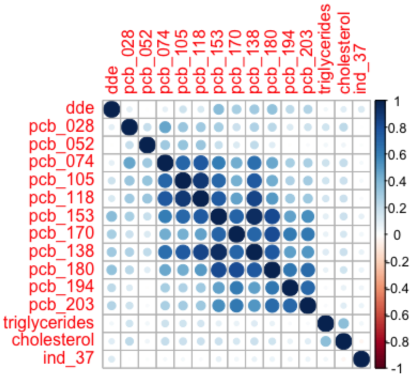
- Data: A study by Longnecker et al. (2001), comprised of 2380 observations of pregnant women.
- Goal: Assess how DDE and PCBs relate to risk of premature delivery.

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

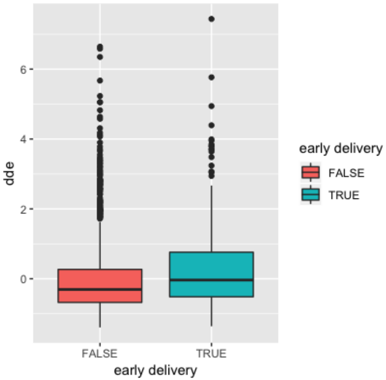
- Possible approaches: Simple sum, PCA, Factor Analysis.

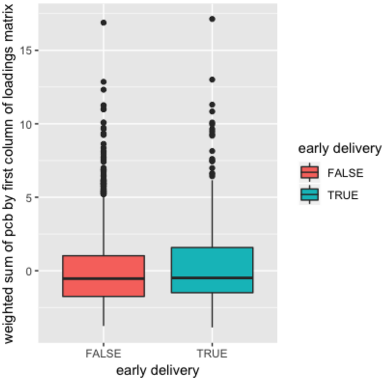
Loadings:

	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

- Binary response: use logistic regression?

Model

- Binary response: use logistic regression?
- Domain knowledge about chemical effects?

Model

- Binary response: use logistic regression?
- Domain knowledge about chemical effects?
- No effect when concentration is below some lower bound.

Model

- Binary response: use logistic regression?
- Domain knowledge about chemical effects?
- No effect when concentration is below some lower bound.
- Constant effect after reaching some upper bound.

Model

- Binary response: use logistic regression?
- Domain knowledge about chemical effects?
- No effect when concentration is below some lower bound.
- Constant effect after reaching some upper bound.
- Generalized Additive Model (GAM)

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 numeric variables: DDE, Principal Components of PCBs (PC1-4), Maternal Age, etc.
- $z_{.k}$s include categorical variables.

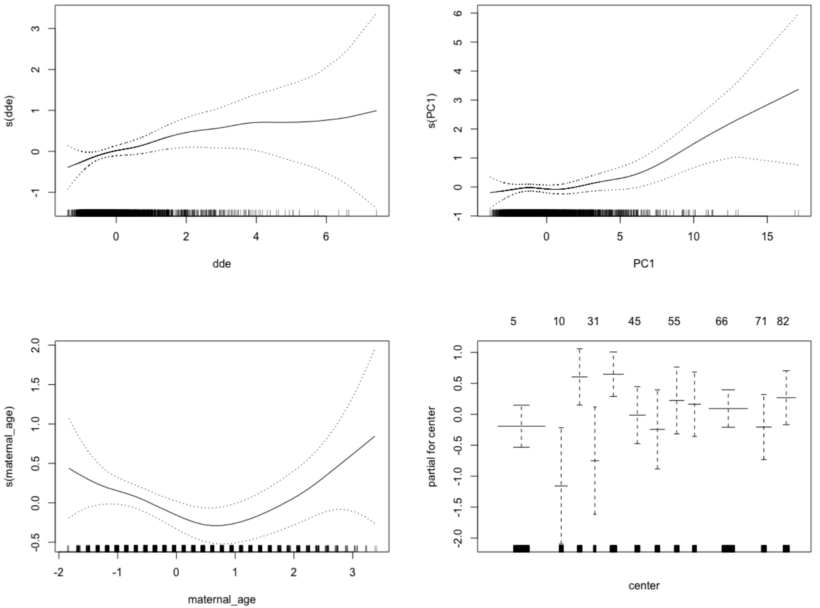
GAM Outputs

```
Anova for Parametric Effects
                    Df  Sum Sq Mean Sq F value    Pr(>F)    
s(dde)               1   31.36 31.3586 30.9091 3.013e-08 ***
s(PC1)               1    4.32  4.3236  4.2616 0.0390938 *  
s(PC2)               1    1.99  1.9909  1.9624 0.1613916    
s(PC3)               1    0.04  0.0412  0.0406 0.8402589    
s(PC4)               1    2.15  2.1509  2.1200 0.1455191    
s(triglycerides)     1    4.27  4.2715  4.2103 0.0402917 *  
score_education      1    7.43  7.4334  7.3268 0.0068429 ** 
score_income         1    8.31  8.3074  8.1883 0.0042538 ** 
score_occupation     1    1.70  1.7046  1.6801 0.1950363    
s(maternal_age)      1    1.67  1.6733  1.6493 0.1991773    
s(cholesterol)       1   10.18 10.1815 10.0356 0.0015554 ** 
smoking_status       1    0.72  0.7212  0.7109 0.3992320    
center              11   35.46  3.2234  3.1772 0.0002718 ***
race                 2    1.42  0.7082  0.6981 0.4976518    
Residuals         2329 2362.87  1.0145                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

GAM Outputs

```
Anova for Nonparametric Effects
                    Npar Df Npar Chisq   P(Chi)
(Intercept)
s(dde)                    3     1.8096 0.612873
s(PC1)                    3     9.8699 0.019707 *
s(PC2)                    3     7.0668 0.069799 .
s(PC3)                    3     2.6365 0.451070
s(PC4)                    3     5.2600 0.153722
s(triglycerides)          3     2.8599 0.413778
score_education
score_income
score_occupation
s(maternal_age)           3    11.4106 0.009702 **
s(cholesterol)            3     2.4425 0.485777
smoking_status
center
race
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GAM Outputs



s(dde)
dde
s(PC1)
PC1
s(maternal_age)
maternal_age
partial for center
center
5
10
31
45
55
66
71
82

Model

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

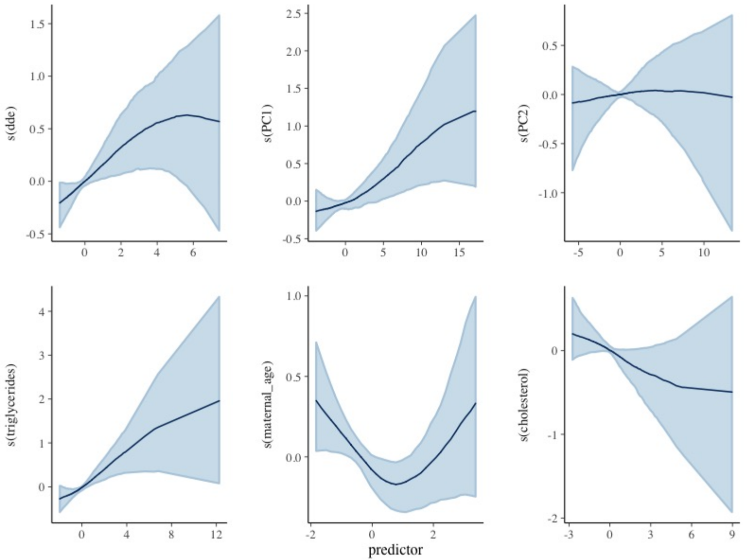
Model

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

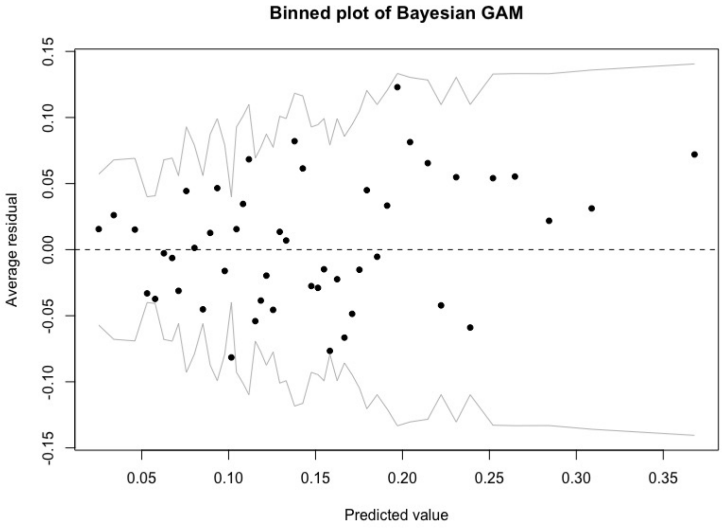
$$g(Y_i) = \beta_0 + \sum_{j=1}^{m} f_j(x_{ij}) + \sum_{k=1}^{l} \beta_k z_{ik}$$

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

Model Results - align with frequentist model



Model Check



Binned plot of Bayesian GAM
Average residual
Predicted value
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
0.05
0.10
0.15
0.20
0.25
0.30
0.35

Discussion

- Deal with different centers

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)

Discussion

- Specialized prior may give narrower credible intervals.
- Including Interactions: Bayesian Factor Analysis (Ferrari, F. and Dunson, D.B. 2019)
- Model with variable selection (e.g. LASSO, GAM + penalty)