Case Study 1: National Collaborative Perinatal Project

Background

The data are taken from the National Collaborative Perinatal Project (CPP). Women were enrolled during pregnancy through different medical centers and then the kids were followed in order to collect both pregnancy and childhood development outcomes. We consider a subsample of 2380 women and children for this analysis, which was studied by [Longnecker et al., 2001]. A particular focus of the Longnecker et al substudy was in assaying serum samples from the original larger study to obtain information on exposures in order to assess the relationship between these exposures to the women and adverse pregnancy and developmental outcomes in their children. Two exposures of particular interest are Dichlorodiphenyldichloroethylene (DDE) and Polychlorinated Biphenyls (PCBs), which are breakdown products in the body of chemicals that have been historically used to treat crops to protect them from predation. These chemicals persist in the environment and are lipophilic, building up in fatty deposits in human tissues. Hence, each of us carries around our own body burden of these chemicals, potentially impacting our health.

The data

The dataset contains demographic variables, such as race, age, and socio-economic index, along with smoking status and concentration doses for DDE and PCBs. In addition, data are available on levels of cholesterol and triglycerides in serum; these variables are relevant since DDE/PCBs are stored in fat and cholesterol/triglycerides provide measurements of the levels of circulating fats (being somewhat informal) in serum.

Goal

The overarching goal of the analysis is to assess how DDE and PCBs relate to risk of premature delivery. Premature delivery is typically defined as a gestational age at delivery of 37 weeks or less, but it is important to note that deliveries occurring right at the cutoff have similar clinical outcomes to full term deliveries, while deliveries occurring substantially less than 37 weeks (early preterm) are associated with substantial risk of short and long term morbidity and mortality. Ideally we would like to infer a causal effect of these exposures on risk of premature deliveries of different severities, while investigating the dose response relationship. However, these data are not collected in a randomized trial but are the result of an observational epidemiology study. Hence, epidemiologists typically focus on assessing associations, while adjusting for covariates that may confound exposure-outcome relationships. In addressing the above interests, it is important to take into account heterogeneity across study centers.

Variable key

gestational_age = gestational age (in weeks)

dde = concentration of dde (ug/dL)

pcb_* = concentration of pcb_* (ng/dL)

albumin = concentration of albumin (g/dL)

cholesterol = concentration of cholesterol (g/dL)

triglycerides = concentration of triglycerides (g/dL)

race

score_education

score_income

score_occupation

maternal_age = age of mother

smoking_status = mother smoking

center

```
###load data & packages
data<-readRDS("~/Downloads/Case-Study-1-master/Longnecker.rds")
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

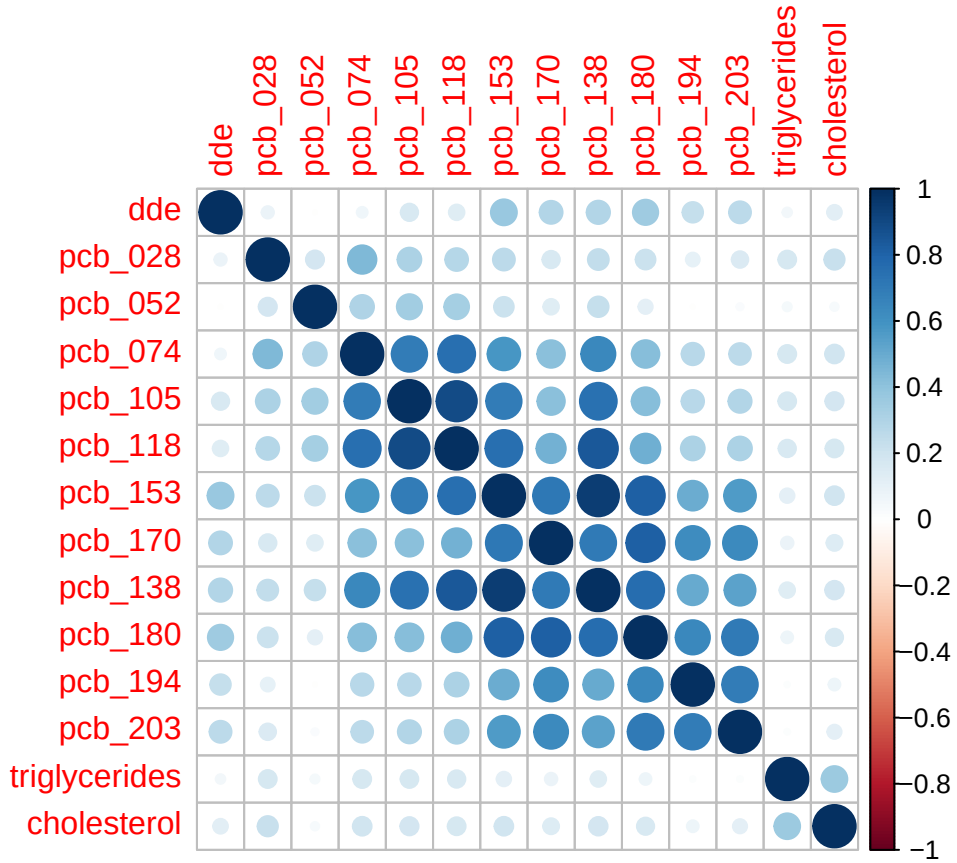
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
X=data %>% select(-albumin,-gestational_age)
y=data %>% select(gestational_age)
predictors=na.omit(X[,1:12])
X=na.omit(X)
```

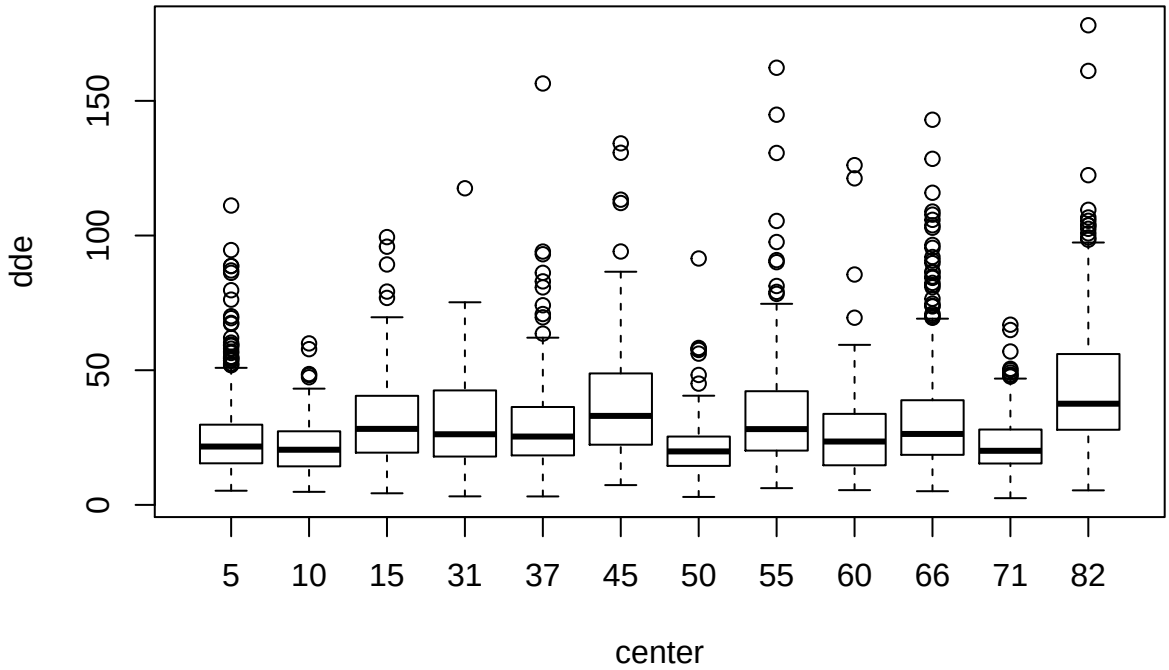
Correlation heat map. There are relatuvely high dependency between pcb variables, which may cause the multicolinearity issue. The association between treatment variables and tri or cho is weak.

```
mea=X %>% select(dde:pcb_203,triglycerides,cholesterol)
corrplot(cor(mea))
```



Boxplot by different centers.Heterogenity exists across centers. May add a random effect for different centers.

```
boxplot(dde~center,data=data)
```



Evalute the marginal association. Very clear marginal association between dde and preterm.

```r
library(ggplot2)
data$preterm=data$gestational_age<37
ggplot(data, aes(x=preterm, y=dde)) +  geom_boxplot()
```