STA 723: Case study 1

January 20, 2020

Professor David Dunson

Olivier Binette, Joe Mathews and Brian Kundinger

Exploratory Data Analysis

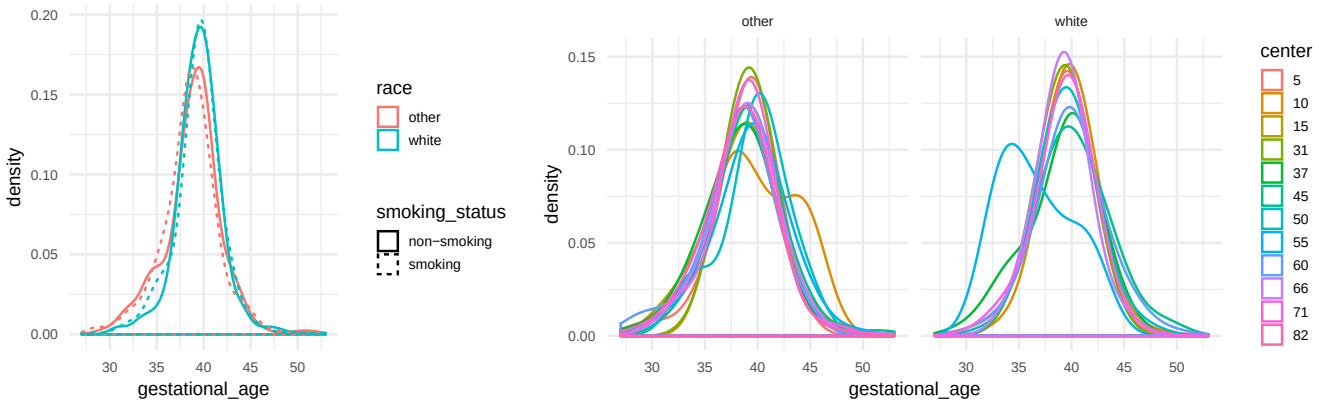
This dataset from the National Collaborative Perinatal Project (CPP) relates gestational age to chemical exposure (DDE and PCBs) and other factors (socio-economic and health-related) in 2380 pregnant women. The goal is to assess how exposure to DDE and PCBs impact the risk of preterm birth, defined as delivery before 37 weeks.

Data cleaning and manipulations

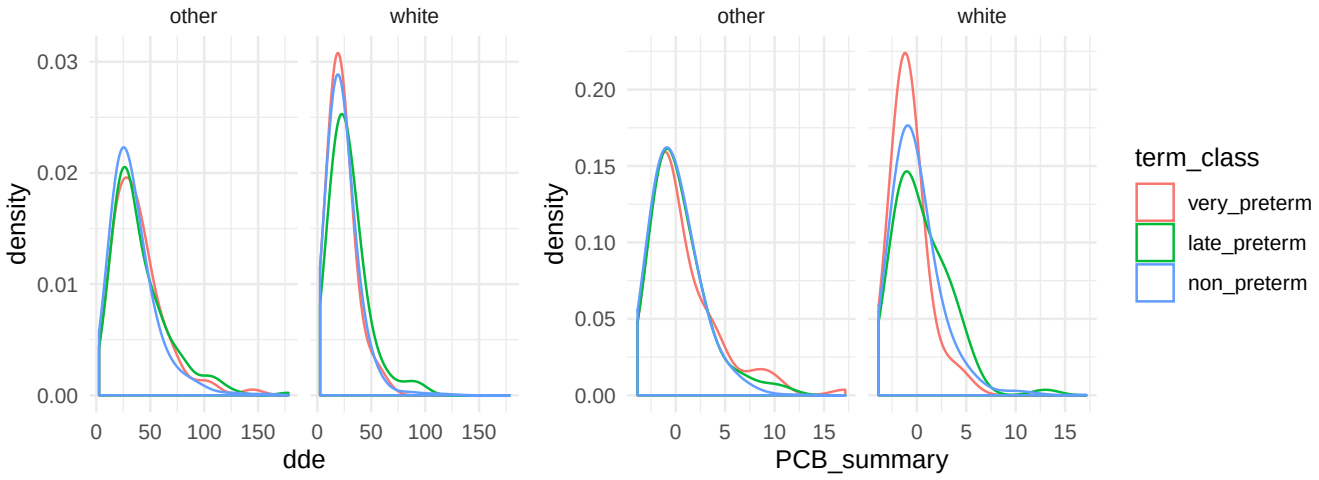
We removed pregnancies over 55 weeks from the dataset. We also dropped the `albumin` variable, which contains 93% missing values. The variables `score_income`, `score_occupation` and `score_education` contain 21% missing values. Otherwise the data contains only one observation (# 1857) with missing $PCBs$ values, which we remove. Given the low sample size for the "other" `race` classification ($n = 123$), we combined "black" and "other" in a single class of size 1336. Finally, to help with visualization and interpretation, we summarized the different PBCs with a positively weighted average. The weights were chosen to minimize the sum of squared orthogonal residuals to normalized PCBs. This ensures that this one-dimensional summary is a relatively good approximation to the PCBs data.

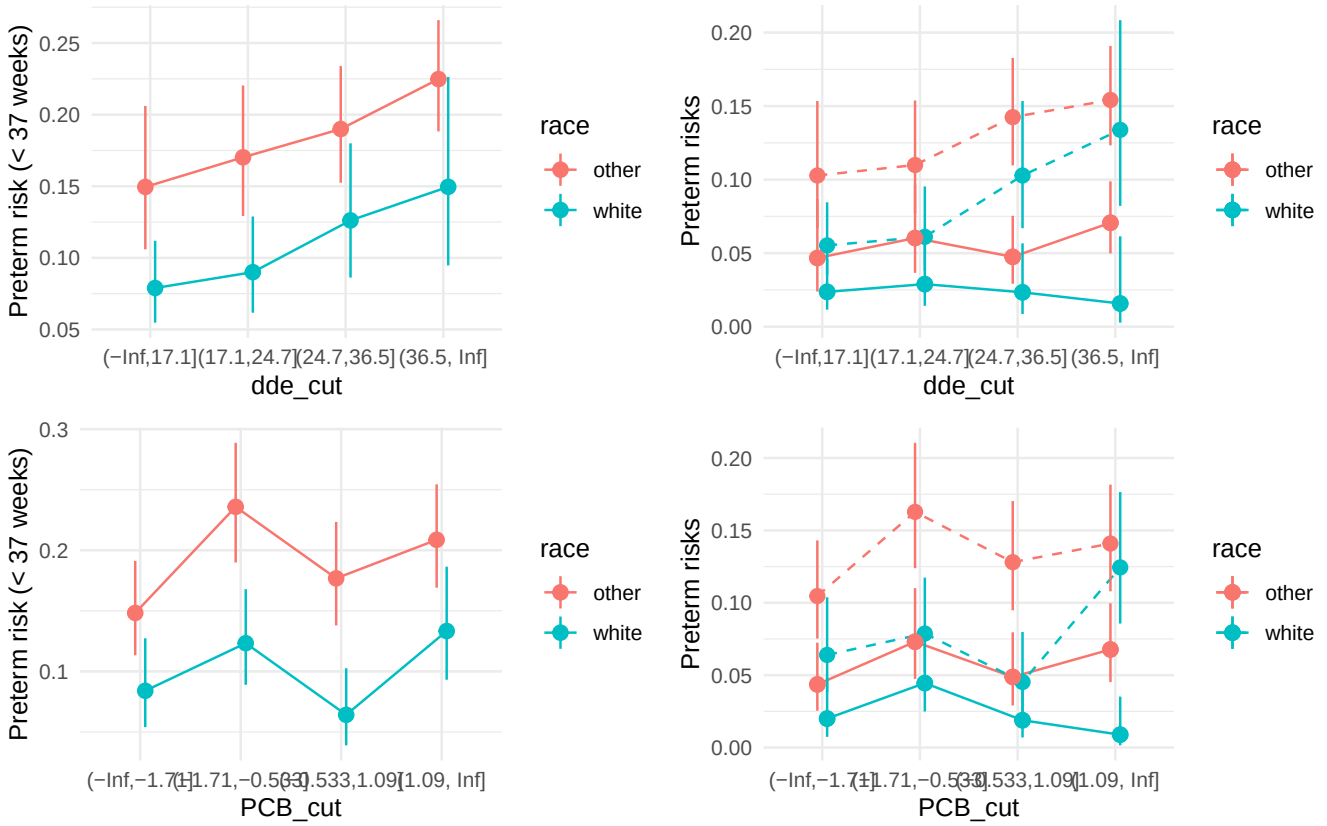
Gestational age

Gestational age distribution by groups.



Relationship with DDE and PCBs: violin plots vs quantiles + preterm risk plot





Colinearity between PCBs, dde and other covariates; partial correlation plot between PCBs and dde.

```
out = complete_data %>%
  glm(preterm ~ smoking_status + score_income + score_education + score_occupation + maternal_age

summary(out)

##
## Call:
## glm(formula = preterm ~ smoking_status + score_income + score_education +
##     score_occupation + maternal_age + cholesterol + dde + PCB_summary +
##     center, data = .)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.44249  -0.17729  -0.11732  -0.05052   1.01360
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.1668236  0.0616809   2.705 0.006902 **
## smoking_statussmoking   0.0217634  0.0167172   1.302 0.193128
## score_income           -0.0002458  0.0003324  -0.740 0.459648
## score_education        -0.0004659  0.0003748  -1.243 0.213977
## score_occupation       -0.0004074  0.0003384  -1.204 0.228699
## maternal_age           -0.0005418  0.0013643  -0.397 0.691348
## cholesterol            -0.0001544  0.0001302  -1.185 0.236018
## dde                     0.0016481  0.0004678   3.523 0.000437 ***
```

```
## PCB_summary            0.0103089  0.0039287   2.624 0.008762 **
## center10              -0.0433844  0.0368928  -1.176 0.239764
## center15               0.1078801  0.0394419   2.735 0.006295 **
## center31              -0.0544777  0.0479512  -1.136 0.256060
## center37               0.1115244  0.0331687   3.362 0.000789 ***
## center45               0.0109522  0.0378092   0.290 0.772101
## center50               0.0131070  0.0364000   0.360 0.718827
## center55               0.0030898  0.0586433   0.053 0.957986
## center60               0.0284479  0.0410994   0.692 0.488916
## center66               0.0200082  0.0291803   0.686 0.493006
## center71              -0.0128947  0.0379423  -0.340 0.734009
## center82               0.0585673  0.0388284   1.508 0.131634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1211502)
##
##     Null deviance: 234.17  on 1851  degrees of freedom
## Residual deviance: 221.95  on 1832  degrees of freedom
## AIC: 1368.6
##
## Number of Fisher Scoring iterations: 2
```

```
outrf = randomForest(gestational_age ~ smoking_status + race + center + score_education + score_ir
```

```
## Error in randomForest(gestational_age ~ smoking_status + race + center + :  could not find
function "randomForest"
```

```
residuals = complete_data$gestational_age - predict(outrf)
```

```
## Error in predict(outrf):  object 'outrf' not found
```

```
cor.test(residuals, complete_data$dde)
```

```
## Error in cor.test.default(residuals, complete_data$dde):  'x' and 'y' must have the same
length
```