cs1Xiaojun_summary

```
library(ggplot2)
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
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
library(monomvn)
```

```
## Loading required package: pls

##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##     loadings

## Loading required package: lars

## Loaded lars 1.2
```

```
library(BAS)
library(R2jags)
```

```
## Loading required package: rjags

## Loading required package: coda

## Linked to JAGS 4.3.0

## Loaded modules: basemod,bugs

##
## Attaching package: 'R2jags'

## The following object is masked from 'package:coda':
##
##     traceplot
```

```
Longnecker <- as.data.frame(readRDS("Longnecker.rds"))
head(Longnecker)
```

```
##     dde pcb_028 pcb_052 pcb_074 pcb_105 pcb_118 pcb_153 pcb_170 pcb_138
## 1 24.56    0.22       0    0.24    0.24    0.85    0.76    0.18    0.81
## 2 15.56    0.20       0    0.22    0.17    0.57    0.69    0.00    0.58
## 3 54.80    0.28       0    0.39    0.09    0.82    1.32    0.33    1.13
## 4 15.00    0.14       0    0.20    0.22    0.73    0.51    0.00    0.60
## 5 33.54    0.17       0    0.21    0.05    0.56    0.75    0.22    0.79
## 6 22.68    0.16       0    0.00    0.12    0.56    0.94    0.00    0.64
##   pcb_180 pcb_194 pcb_203 albumin triglycerides  race score_education
## 1    0.38    0.13    0.11      NA           294 white              97
## 2    0.26    0.00    0.00       3           180 white              15
## 3    0.61    0.12    0.14      NA           278 white              91
## 4    0.22    0.00    0.00      NA           182 white              NA
## 5    0.42    0.12    0.15      NA           201 white              43
## 6    0.42    0.08    0.13      NA           326 white              64
##   score_income score_occupation maternal_age smoking_status cholesterol
## 1           65               94           27              1         385
## 2           78               35           28              1          93
## 3           94               84           31              0         273
## 4           NA               NA           33              0         166
## 5           43                5           28              1         275
## 6           91               94           27              0         319
##   gestational_age center
## 1              41      5
## 2              41      5
## 3              36      5
## 4              40      5
## 5              40      5
## 6              40      5
```

```
Longnecker$center<- as.factor(as.character(Longnecker$center))
Longnecker$center <- relevel(Longnecker$center, ref = "5")
Longnecker[Longnecker$gestational_age > 46, "gestational_age"] <- 46 # Truncate at age == 46
```

```r
# Complete case analysis for missing variables
x <- model.matrix(gestational_age ~ . - center, data = Longnecker)
miss_indices <- as.integer(row.names(x))
Longnecker_complete <- Longnecker[miss_indices,] # Only those non-missing obs.
data_age_na<- Longnecker_complete
data_age<- Longnecker_complete
```

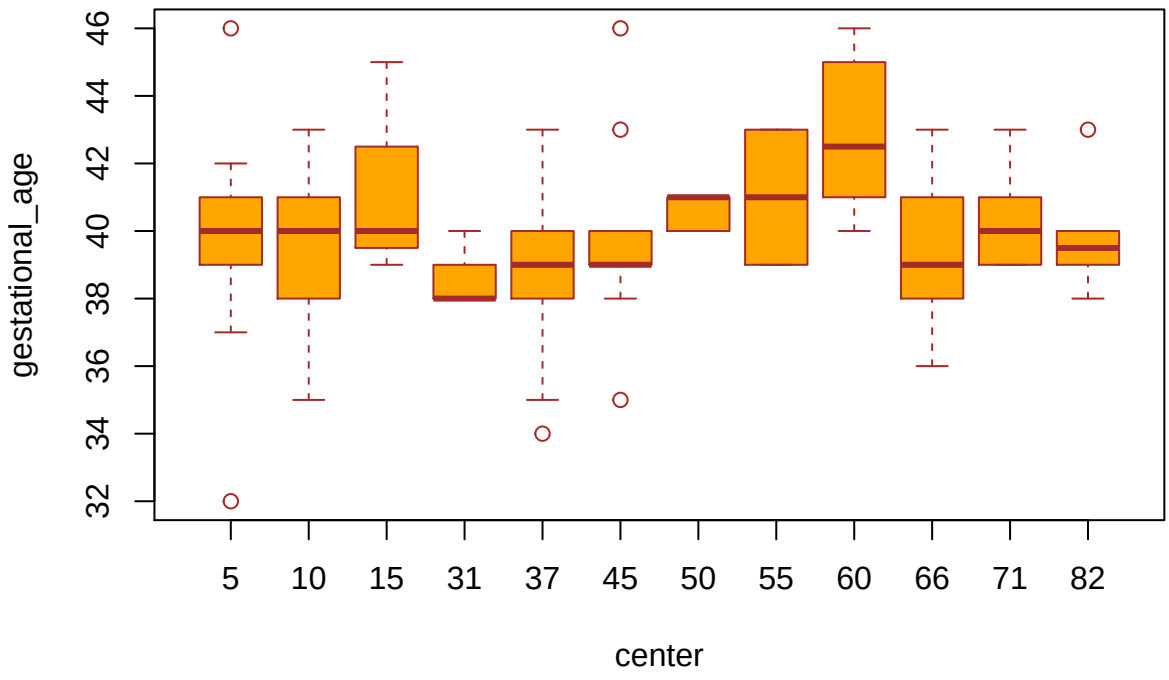
Some EDA that Yunran did not talk about

Center ID 5 has the most observations (481) across all centers, and center ID 31 has the fewet observations (78). If we consider 37 weeks or less as premature, then center ID 15 and 37 would have the highest premature rate, which is approximately 25%. We suspect that it might be because the black dominates these centers. However, center ID 45 and 66 also consist of mainly black people, but their premature rate is not as high as the center ID 15 and 37. Thus, we guess that there might be heterogeneity between centers. Also, we observed that there are unbalanced number of white and black if the maternal age is lower than 20, and center ID 5 and 66 have more older mother than the others. Finally, if we look at the density plot for gestational age grouped by race, we can tell that white would have less risk for premature comparing to the black and the other.

```r
table(data_age$center)
```

```
##
##  5 10 15 31 37 45 50 55 60 66 71 82
## 38  8  4  4 14 10  5  2 10  9  9  6
```

```r
boxplot(gestational_age ~ center, data=data_age, col="orange", border="brown")
```



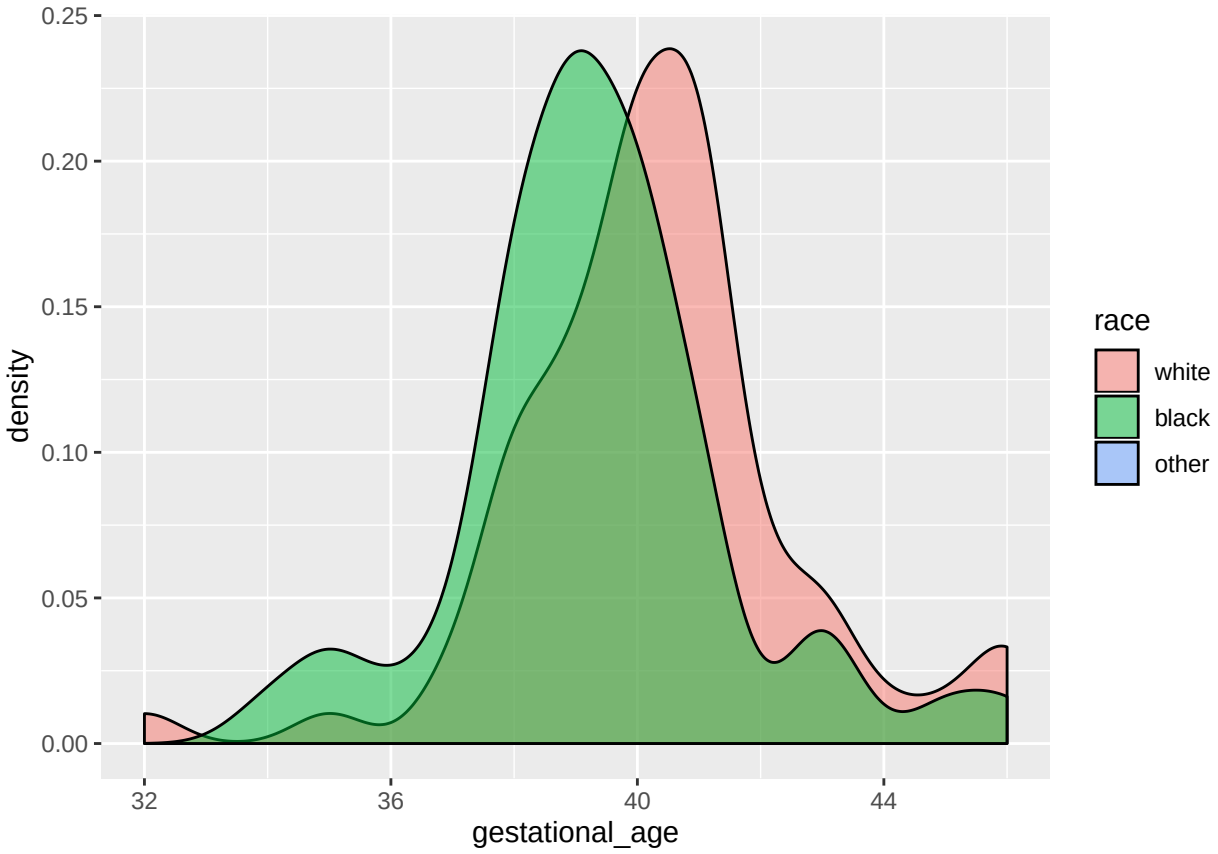

```
premature<- c()
premature_ratio<- c()
for (i in unique(data_age$center)){
   pre <- sum(data_age[data_age$center == i, ]$gestational_age <37)
   premature<- c(premature, pre)
   pre_ratio<- pre/sum(data_age$center == i)
   premature_ratio<- c(premature_ratio, pre_ratio)
}

data.frame(ratio= premature_ratio, center= unique(data_age$center))
```

```
##         ratio center
## 1  0.02631579      5
## 2  0.12500000     10
## 3  0.00000000     15
## 4  0.00000000     31
## 5  0.14285714     37
## 6  0.10000000     45
## 7  0.00000000     50
## 8  0.00000000     60
## 9  0.11111111     66
## 10 0.00000000     71
## 11 0.00000000     82
## 12 0.00000000     55
```

```
ggplot(data_age, aes(gestational_age, fill = race)) + geom_density(alpha = 0.5)
```

```
## Warning: Groups with fewer than two data points have been dropped.
```




```
table(data_age$center, data_age$race)
```

```
##
##       white black other
##   5      35     3     0
##   10      8     0     0
##   15      0     4     0
##   31      0     4     0
##   37      3    11     0
##   45      2     8     0
##   50      5     0     0
##   55      0     1     1
##   60      9     1     0
##   66      0     9     0
##   71      5     4     0
##   82      0     6     0
```

```
table(data_age$maternal_age, data_age$race) ## unbalanced number between white and black if the materna
```

```
##
##      white black other
##   14     0     2     0
##   15     0     1     0
##   16     1     5     0
##   17     2     1     0
##   18     4     4     0
##   19     4     5     0
##   20    10     1     0
##   21     6     2     0
##   22     5     4     0
##   23     4     2     0
##   24     4     2     0
##   25     2     0     0
##   26     3     1     0
##   27     4     5     0
##   28     3     3     0
##   29     2     2     0
##   30     0     0     1
##   31     1     1     0
##   32     0     3     0
##   33     0     4     0
##   34     3     0     0
##   35     2     0     0
##   36     1     1     0
##   37     1     0     0
##   38     0     1     0
##   39     1     1     0
##   41     2     0     0
##   42     1     0     0
##   45     1     0     0
```

```
table(data_age$maternal_age, data_age$center) ### center 5 and 66 have more older mother
```

```
##
##      5 10 15 31 37 45 50 55 60 66 71 82
##   14 0  0  0  0  1  1  0  0  0  0  0  0
##   15 0  0  0  0  0  0  0  0  0  0  0  1
##   16 0  0  1  0  0  2  0  0  0  3  0  0
##   17 0  0  0  0  0  0  0  0  2  0  1  0
##   18 2  0  2  0  1  0  0  0  1  1  0  1
##   19 2  0  0  0  2  0  1  0  1  2  0  1
##   20 6  2  0  0  1  0  1  0  0  0  1  0
##   21 4  0  0  0  0  1  0  0  2  0  1  0
##   22 4  0  0  0  1  1  1  0  0  1  0  1
##   23 0  1  0  1  0  0  0  0  1  1  2  0
##   24 1  0  0  0  2  0  2  0  1  0  0  0
##   25 1  1  0  0  0  0  0  0  0  0  0  0
##   26 1  0  0  0  1  1  0  0  1  0  0  0
##   27 3  1  0  0  1  1  0  0  0  0  2  1
##   28 2  0  0  0  1  1  0  0  1  0  1  0
##   29 2  0  0  0  1  0  0  0  0  1  0  0
##   30 0  0  0  0  0  0  0  1  0  0  0  0
##   31 1  0  0  0  0  1  0  0  0  0  0  0
```

```
##   32 1  0  0  0  1  1  0  0  0  0  0  0
##   33 0  0  0  2  1  0  0  0  0  0  0  1
##   34 1  2  0  0  0  0  0  0  0  0  0  0
##   35 2  0  0  0  0  0  0  0  0  0  0  0
##   36 1  0  0  1  0  0  0  0  0  0  0  0
##   37 0  0  0  0  0  0  0  0  0  0  1  0
##   38 0  0  0  0  0  0  0  1  0  0  0  0
##   39 1  0  1  0  0  0  0  0  0  0  0  0
##   41 1  1  0  0  0  0  0  0  0  0  0  0
##   42 1  0  0  0  0  0  0  0  0  0  0  0
##   45 1  0  0  0  0  0  0  0  0  0  0  0
```

Bayesian Model Averaging

We first started with a full model with all main effects, and interactions between race and all demographic variables as well as center, and then perform BMA via bas.lm. However, the variables that are included in the best model are only race and triglycerides. Since most main effects do not show any significance associated with gestational age, we decided to remove all interaction terms. Again, no extra covariates appear in the best model. Then we decided to remove some PCBs. On the one hand, PCBs are correlated, and including all of them is not a wise choice. One the other hand, some of the PCBs are zero-inflated, which would be a problem for modeling. Thus, we removed the PCBs which have a bunch of 0s, such as PCB_105, PCB_138, etc.. Even though we used a much simplier model, we still don't see any extra covariates introduced into the best model, so we would like to use other techniques to do model selection, and think about more how to deal with PCBs.

```
age.bas1 = bas.lm(gestational_age~ dde + pcb_028 + pcb_052 + pcb_074 + pcb_105 + pcb_118 + pcb_153+ pcb
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored
```

```
summary(age.bas1)
```

```
##                          P(B != 0 | Y)  model 1   model 2   model 3
## Intercept                  1.000000000 1.000000 1.0000000 1.0000000
## dde                        0.344709420 0.000000 0.0000000 0.0000000
## pcb_028                    0.194192501 0.000000 0.0000000 0.0000000
## pcb_052                    0.247739331 0.000000 0.0000000 0.0000000
## pcb_074                    0.290811386 0.000000 0.0000000 0.0000000
## pcb_105                    0.298360589 0.000000 0.0000000 0.0000000
## pcb_118                    0.241543802 0.000000 0.0000000 0.0000000
## pcb_153                    0.264655137 0.000000 0.0000000 0.0000000
## pcb_170                    0.206143701 0.000000 0.0000000 0.0000000
## pcb_138                    0.242205295 0.000000 0.0000000 0.0000000
## pcb_180                    0.240367435 0.000000 0.0000000 0.0000000
## pcb_194                    0.957118891 1.000000 1.0000000 1.0000000
## pcb_203                    0.359480489 0.000000 0.0000000 0.0000000
## raceblack                  0.622406880 0.000000 0.0000000 0.0000000
## raceother                  0.622406880 0.000000 0.0000000 0.0000000
## triglycerides              0.680239774 0.000000 0.0000000 1.0000000
## cholesterol                0.281318621 0.000000 0.0000000 0.0000000
## maternal_age               0.358406416 0.000000 0.0000000 0.0000000
## smoking_status             0.548094628 0.000000 1.0000000 0.0000000
```

```
## score_education            0.479200480 0.000000 0.0000000 0.0000000
## score_income               0.212461981 0.000000 0.0000000 0.0000000
## score_occupation           0.263758371 0.000000 0.0000000 0.0000000
## center10                   0.017191997 0.000000 0.0000000 0.0000000
## center15                   0.017191997 0.000000 0.0000000 0.0000000
## center31                   0.017191997 0.000000 0.0000000 0.0000000
## center37                   0.017191997 0.000000 0.0000000 0.0000000
## center45                   0.017191997 0.000000 0.0000000 0.0000000
## center50                   0.017191997 0.000000 0.0000000 0.0000000
## center55                   0.017191997 0.000000 0.0000000 0.0000000
## center60                   0.017191997 0.000000 0.0000000 0.0000000
## center66                   0.017191997 0.000000 0.0000000 0.0000000
## center71                   0.017191997 0.000000 0.0000000 0.0000000
## center82                   0.017191997 0.000000 0.0000000 0.0000000
## dde:raceblack              0.097334934 0.000000 0.0000000 0.0000000
## dde:raceother              0.097334934 0.000000 0.0000000 0.0000000
## raceblack:triglycerides    0.220768150 0.000000 0.0000000 0.0000000
## raceother:triglycerides    0.220768150 0.000000 0.0000000 0.0000000
## raceblack:cholesterol      0.042748128 0.000000 0.0000000 0.0000000
## raceother:cholesterol      0.042748128 0.000000 0.0000000 0.0000000
## raceblack:maternal_age     0.063448765 0.000000 0.0000000 0.0000000
## raceother:maternal_age     0.063448765 0.000000 0.0000000 0.0000000
## raceblack:smoking_status   0.113981369 0.000000 0.0000000 0.0000000
## raceother:smoking_status   0.113981369 0.000000 0.0000000 0.0000000
## raceblack:center10         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center10         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center15         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center15         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center31         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center31         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center37         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center37         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center45         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center45         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center50         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center50         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center55         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center55         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center60         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center60         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center66         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center66         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center71         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center71         0.000576249 0.000000 0.0000000 0.0000000
## raceblack:center82         0.000576249 0.000000 0.0000000 0.0000000
## raceother:center82         0.000576249 0.000000 0.0000000 0.0000000
## BF                                    NA 1.000000 0.4806207 0.3365299
## PostProbs                             NA 0.002100 0.0019000 0.0019000
## R2                                    NA 0.123600 0.1463000 0.1410000
## dim                                   NA 2.000000 3.0000000 3.0000000
## logmarg                               NA 5.034738 4.3020610 3.9456696
##                            model 4   model 5
## Intercept                1.0000000 1.0000000
## dde                      0.0000000 0.0000000
```

```
## pcb_028                 0.0000000 0.0000000
## pcb_052                 0.0000000 1.0000000
## pcb_074                 0.0000000 0.0000000
## pcb_105                 0.0000000 0.0000000
## pcb_118                 0.0000000 0.0000000
## pcb_153                 0.0000000 0.0000000
## pcb_170                 0.0000000 0.0000000
## pcb_138                 0.0000000 0.0000000
## pcb_180                 0.0000000 0.0000000
## pcb_194                 1.0000000 1.0000000
## pcb_203                 0.0000000 0.0000000
## raceblack               0.0000000 0.0000000
## raceother               0.0000000 0.0000000
## triglycerides           0.0000000 0.0000000
## cholesterol             0.0000000 0.0000000
## maternal_age            1.0000000 0.0000000
## smoking_status          0.0000000 0.0000000
## score_education         0.0000000 0.0000000
## score_income            0.0000000 0.0000000
## score_occupation        0.0000000 0.0000000
## center10                0.0000000 0.0000000
## center15                0.0000000 0.0000000
## center31                0.0000000 0.0000000
## center37                0.0000000 0.0000000
## center45                0.0000000 0.0000000
## center50                0.0000000 0.0000000
## center55                0.0000000 0.0000000
## center60                0.0000000 0.0000000
## center66                0.0000000 0.0000000
## center71                0.0000000 0.0000000
## center82                0.0000000 0.0000000
## dde:raceblack           0.0000000 0.0000000
## dde:raceother           0.0000000 0.0000000
## raceblack:triglycerides  0.0000000 0.0000000
## raceother:triglycerides  0.0000000 0.0000000
## raceblack:cholesterol    0.0000000 0.0000000
## raceother:cholesterol    0.0000000 0.0000000
## raceblack:maternal_age   0.0000000 0.0000000
## raceother:maternal_age   0.0000000 0.0000000
## raceblack:smoking_status 0.0000000 0.0000000
## raceother:smoking_status 0.0000000 0.0000000
## raceblack:center10       0.0000000 0.0000000
## raceother:center10       0.0000000 0.0000000
## raceblack:center15       0.0000000 0.0000000
## raceother:center15       0.0000000 0.0000000
## raceblack:center31       0.0000000 0.0000000
## raceother:center31       0.0000000 0.0000000
## raceblack:center37       0.0000000 0.0000000
## raceother:center37       0.0000000 0.0000000
## raceblack:center45       0.0000000 0.0000000
## raceother:center45       0.0000000 0.0000000
## raceblack:center50       0.0000000 0.0000000
## raceother:center50       0.0000000 0.0000000
## raceblack:center55       0.0000000 0.0000000
```

```
## raceother:center55       0.0000000 0.0000000
## raceblack:center60       0.0000000 0.0000000
## raceother:center60       0.0000000 0.0000000
## raceblack:center66       0.0000000 0.0000000
## raceother:center66       0.0000000 0.0000000
## raceblack:center71       0.0000000 0.0000000
## raceother:center71       0.0000000 0.0000000
## raceblack:center82       0.0000000 0.0000000
## raceother:center82       0.0000000 0.0000000
## BF                       0.2123928 0.1614827
## PostProbs                0.0012000 0.0010000
## R2                       0.1341000 0.1299000
## dim                      3.0000000 3.0000000
## logmarg                  3.4854200 3.2113809
```

```
age.bas2 = bas.lm(gestational_age~ dde + pcb_028 + pcb_052 + pcb_074 + pcb_105 + pcb_118 + pcb_153+ pcb
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored
```

```
summary(age.bas2)
```

```
##                  P(B != 0 | Y)  model 1   model 2   model 3   model 4
## Intercept          1.000000000 1.000000 1.0000000 1.0000000 1.0000000
## dde                0.191860986 0.000000 0.0000000 0.0000000 0.0000000
## pcb_028            0.165867846 0.000000 0.0000000 0.0000000 0.0000000
## pcb_052            0.205446828 0.000000 0.0000000 0.0000000 0.0000000
## pcb_074            0.248832997 0.000000 0.0000000 0.0000000 0.0000000
## pcb_105            0.276099725 0.000000 0.0000000 0.0000000 0.0000000
## pcb_118            0.211472610 0.000000 0.0000000 0.0000000 0.0000000
## pcb_153            0.240504439 0.000000 0.0000000 0.0000000 0.0000000
## pcb_170            0.183679067 0.000000 0.0000000 0.0000000 0.0000000
## pcb_138            0.201726540 0.000000 0.0000000 0.0000000 0.0000000
## pcb_180            0.225737337 0.000000 0.0000000 0.0000000 0.0000000
## pcb_194            0.942655960 1.000000 1.0000000 1.0000000 1.0000000
## pcb_203            0.256786811 0.000000 0.0000000 0.0000000 0.0000000
## raceblack          0.259410022 0.000000 0.0000000 0.0000000 0.0000000
## raceother          0.259410022 0.000000 0.0000000 0.0000000 0.0000000
## triglycerides      0.405516889 0.000000 0.0000000 1.0000000 0.0000000
## cholesterol        0.178166522 0.000000 0.0000000 0.0000000 0.0000000
## maternal_age       0.234970170 0.000000 0.0000000 0.0000000 0.0000000
## smoking_status     0.380064141 0.000000 1.0000000 0.0000000 0.0000000
## score_education    0.357636503 0.000000 0.0000000 0.0000000 1.0000000
## score_income       0.171497158 0.000000 0.0000000 0.0000000 0.0000000
## score_occupation   0.203377588 0.000000 0.0000000 0.0000000 0.0000000
## center10           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center15           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center31           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center37           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center45           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center50           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center55           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center60           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
```

```
## center66           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center71           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## center82           0.004228094 0.000000 0.0000000 0.0000000 0.0000000
## BF                          NA 1.000000 0.4806207 0.3365299 0.1991033
## PostProbs                   NA 0.017000 0.0079000 0.0061000 0.0036000
## R2                          NA 0.123600 0.1463000 0.1410000 0.1331000
## dim                         NA 2.000000 3.0000000 3.0000000 3.0000000
## logmarg                     NA 5.034738 4.3020610 3.9456696 3.4208064
##                      model 5
## Intercept          1.0000000
## dde                0.0000000
## pcb_028            0.0000000
## pcb_052            0.0000000
## pcb_074            0.0000000
## pcb_105            0.0000000
## pcb_118            0.0000000
## pcb_153            0.0000000
## pcb_170            0.0000000
## pcb_138            0.0000000
## pcb_180            0.0000000
## pcb_194            1.0000000
## pcb_203            0.0000000
## raceblack          0.0000000
## raceother          0.0000000
## triglycerides      0.0000000
## cholesterol        0.0000000
## maternal_age       1.0000000
## smoking_status     0.0000000
## score_education    0.0000000
## score_income       0.0000000
## score_occupation   0.0000000
## center10           0.0000000
## center15           0.0000000
## center31           0.0000000
## center37           0.0000000
## center45           0.0000000
## center50           0.0000000
## center55           0.0000000
## center60           0.0000000
## center66           0.0000000
## center71           0.0000000
## center82           0.0000000
## BF                 0.2123928
## PostProbs          0.0035000
## R2                 0.1341000
## dim                3.0000000
## logmarg            3.4854200
```

```
age.bas3 = bas.lm(gestational_age~ dde + pcb_028 + pcb_052 + pcb_074 + pcb_153+ pcb_138 + pcb_180 + pcb
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored
```

```
summary(age.bas3)
```

```
##                  P(B != 0 | Y)  model 1   model 2   model 3   model 4
## Intercept            1.000000 1.000000 1.0000000 1.0000000 1.0000000
## dde                  0.181980 0.000000 0.0000000 0.0000000 0.0000000
## pcb_028              0.159642 0.000000 0.0000000 0.0000000 0.0000000
## pcb_052              0.208118 0.000000 0.0000000 0.0000000 0.0000000
## pcb_074              0.261650 0.000000 0.0000000 0.0000000 0.0000000
## pcb_153              0.199530 0.000000 0.0000000 0.0000000 0.0000000
## pcb_138              0.185968 0.000000 0.0000000 0.0000000 0.0000000
## pcb_180              0.218552 0.000000 0.0000000 0.0000000 0.0000000
## pcb_194              0.957624 1.000000 1.0000000 1.0000000 1.0000000
## pcb_203              0.233948 0.000000 0.0000000 0.0000000 0.0000000
## raceblack            0.232470 0.000000 0.0000000 0.0000000 0.0000000
## raceother            0.232470 0.000000 0.0000000 0.0000000 0.0000000
## triglycerides        0.353090 0.000000 0.0000000 1.0000000 0.0000000
## cholesterol          0.161298 0.000000 0.0000000 0.0000000 0.0000000
## maternal_age         0.224004 0.000000 0.0000000 0.0000000 0.0000000
## smoking_status       0.398684 0.000000 1.0000000 0.0000000 0.0000000
## score_education      0.325498 0.000000 0.0000000 0.0000000 1.0000000
## score_income         0.163352 0.000000 0.0000000 0.0000000 0.0000000
## score_occupation     0.193616 0.000000 0.0000000 0.0000000 0.0000000
## center10             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center15             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center31             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center37             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center45             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center50             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center55             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center60             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center66             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center71             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## center82             0.002386 0.000000 0.0000000 0.0000000 0.0000000
## BF                         NA 1.000000 0.4806207 0.3365299 0.1991033
## PostProbs                  NA 0.032500 0.0171000 0.0117000 0.0078000
## R2                         NA 0.123600 0.1463000 0.1410000 0.1331000
## dim                        NA 2.000000 3.0000000 3.0000000 3.0000000
## logmarg                    NA 5.034738 4.3020610 3.9456696 3.4208064
##                    model 5
## Intercept        1.0000000
## dde              0.0000000
## pcb_028          0.0000000
## pcb_052          0.0000000
## pcb_074          0.0000000
## pcb_153          0.0000000
## pcb_138          0.0000000
## pcb_180          0.0000000
## pcb_194          1.0000000
## pcb_203          0.0000000
## raceblack        0.0000000
## raceother        0.0000000
## triglycerides    0.0000000
## cholesterol      0.0000000
```

```
## maternal_age     1.0000000
## smoking_status   0.0000000
## score_education  0.0000000
## score_income     0.0000000
## score_occupation 0.0000000
## center10         0.0000000
## center15         0.0000000
## center31         0.0000000
## center37         0.0000000
## center45         0.0000000
## center50         0.0000000
## center55         0.0000000
## center60         0.0000000
## center66         0.0000000
## center71         0.0000000
## center82         0.0000000
## BF               0.2123928
## PostProbs        0.0078000
## R2               0.1341000
## dim              3.0000000
## logmarg          3.4854200
```

Horseshoe prior

We applied Horseshoe prior for all the parameters in this problem. The reason is that Horseshoe prior has the following two properties: sparsity and unbiasedness. This allows us to shrink small coefficients towards 0 and meanwhile keeping large coefficients in the model. We adapted the full model with all main effects excluding the PCBs with zero inflation.

```
full.lm<- lm(gestational_age~ dde + pcb_028 + pcb_052 + pcb_074 + pcb_153+ pcb_138 + pcb_180 + pcb_194 +

X = model.matrix(full.lm)
Y = data_age_na$gestational_age
nrow(X); length(Y);
```

```
## [1] 119
```

```
## [1] 119
```

```
bhs.fit = bhs(X[,-1], Y, T=5000, normalize=T)
```

```
## t=100, m=15
## t=200, m=14
## t=300, m=14
## t=400, m=15
## t=500, m=18
## t=600, m=15
## t=700, m=16
## t=800, m=16
## t=900, m=17
## t=1000, m=16
## t=1100, m=13
```

```
## t=1200, m=15
## t=1300, m=18
## t=1400, m=15
## t=1500, m=12
## t=1600, m=13
## t=1700, m=14
## t=1800, m=16
## t=1900, m=16
## t=2000, m=17
## t=2100, m=13
## t=2200, m=15
## t=2300, m=15
## t=2400, m=10
## t=2500, m=15
## t=2600, m=11
## t=2700, m=19
## t=2800, m=17
## t=2900, m=13
## t=3000, m=13
## t=3100, m=17
## t=3200, m=14
## t=3300, m=17
## t=3400, m=20
## t=3500, m=17
## t=3600, m=17
## t=3700, m=12
## t=3800, m=17
## t=3900, m=16
## t=4000, m=13
## t=4100, m=15
## t=4200, m=12
## t=4300, m=15
## t=4400, m=12
## t=4500, m=14
## t=4600, m=15
## t=4700, m=14
## t=4800, m=14
## t=4900, m=18
```

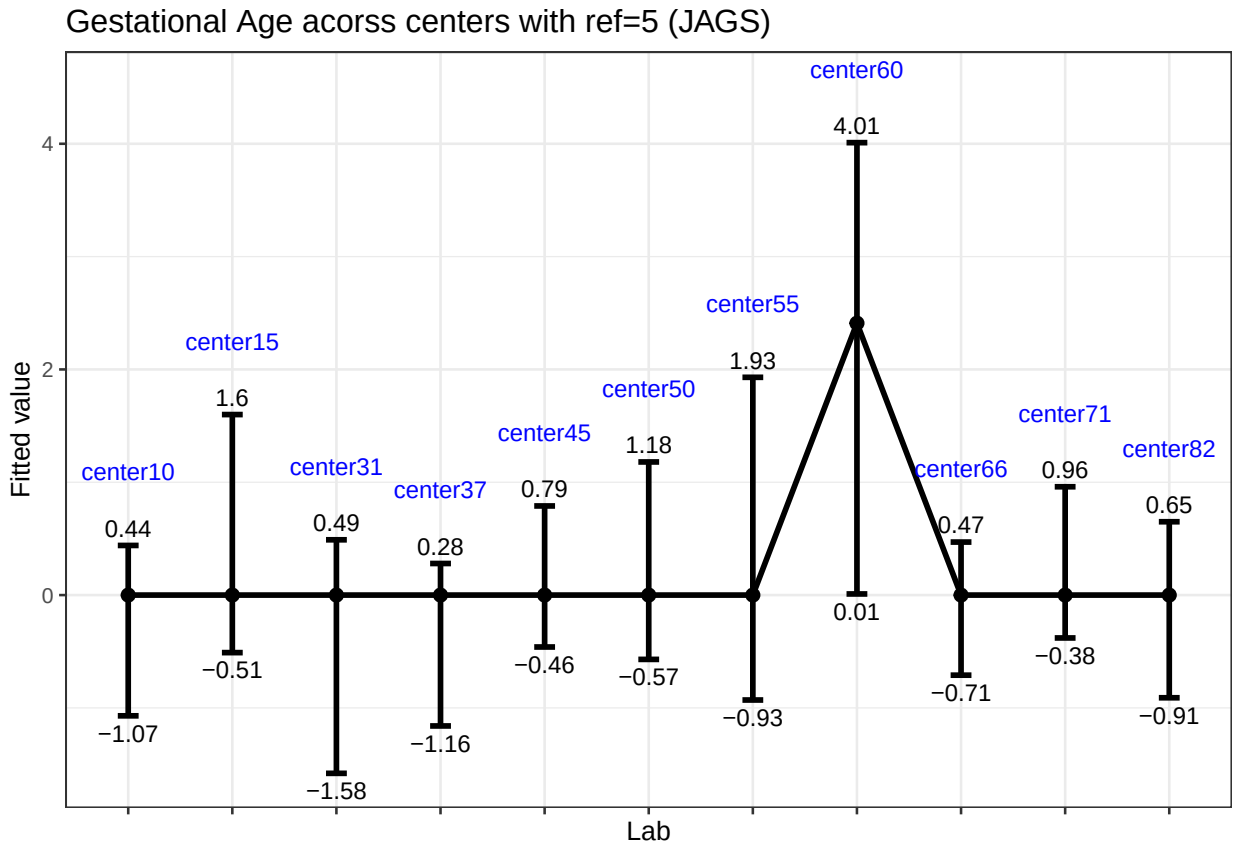
```r
beta.sim = bhs.fit$beta
colnames(beta.sim) = colnames(X)[-1]
quant5 = function(x) {round(quantile(x, c(0.025, 0.5, 0.975)),2)} ## 95% CI

coefs_for_table<- apply(beta.sim, 2, quant5)
knitr::kable(coefs_for_table, format = "latex")
```

	dde	pcb_028	pcb_052	pcb_074	pcb_153	pcb_138	pcb_180	pcb_194	pcb_203	raceblack	ra
2.5%	-0.02	-1.01	-0.76	-0.32	-0.47	-0.39	-2.78	-16.08	-4.48	-0.86	
50%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-6.16	0.00	0.00	
97.5%	0.01	1.64	5.45	1.49	0.48	0.55	1.09	0.17	2.59	0.16	

```r
center = dplyr::select(data.frame(beta.sim),starts_with("center"))
center_fit<- cbind(lwr= apply(center, 2, quant5)[1,], fit= apply(center, 2, quant5)[2,], upr=apply(cent
```

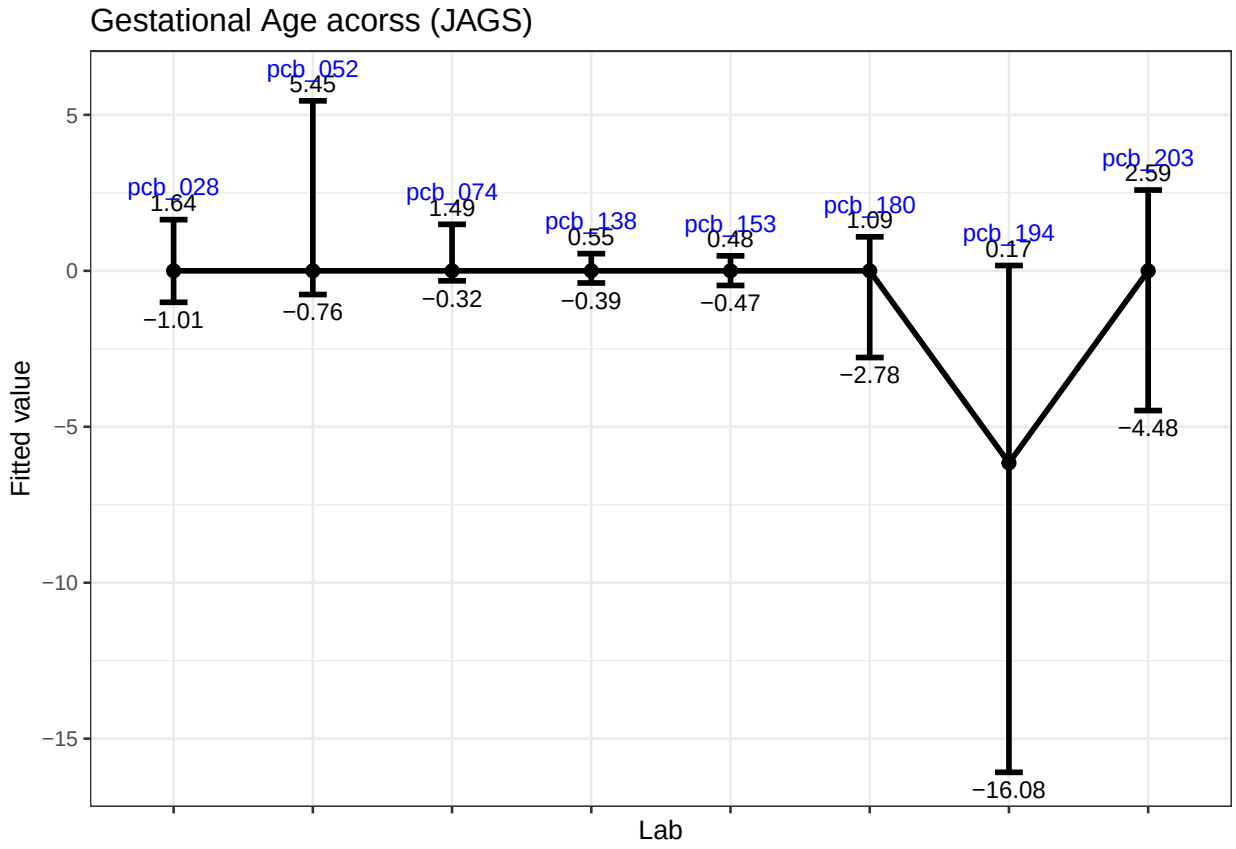
```
ggplot(as.data.frame(center_fit), aes(rownames(center_fit),
  fit, size=10, group=1, ylim=max(upr)+0.8)) +
  theme_bw(base_size=10)+
  geom_point(size=2)+
  geom_line(size=1)+
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.2, size=1)+
  xlab("Lab")+
  ylab("Fitted value")+
  ggtitle("Gestational Age acorss centers with ref=5 (JAGS)")+
  theme(axis.text.x=element_blank())+
  geom_text(aes(label = round(lwr,2), y = lwr), vjust = 1.5, size=3) +
  geom_text(aes(label = round(upr,2), y = upr), vjust = -0.5, size=3) +
  geom_text(aes(label = rownames(center_fit), y = upr+0.5), vjust = -0.5, size=3, col="blue")
```




```
pcb = dplyr::select(data.frame(beta.sim),starts_with("pcb_"))
pcb_fit<- cbind(lwr= apply(pcb, 2, quant5)[1,], fit= apply(pcb, 2, quant5)[2,], upr=apply(pcb, 2, quant5

ggplot(as.data.frame(pcb_fit), aes(rownames(pcb_fit),
  fit, size=10, group=1, ylim=max(upr)+0.8)) +
  theme_bw(base_size=10)+
  geom_point(size=2)+
  geom_line(size=1)+
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.2, size=1)+
  xlab("Lab")+
  ylab("Fitted value")+
```

```
  ggtitle("Gestational Age acorss (JAGS)")+
  theme(axis.text.x=element_blank())+
  geom_text(aes(label = round(lwr,2), y = lwr), vjust = 1.5, size=3) +
  geom_text(aes(label = round(upr,2), y = upr), vjust = -0.5, size=3) +
  geom_text(aes(label = rownames(pcb_fit), y = upr+0.5), vjust = -0.5, size=3, col="blue")
```



Hierarchical Prior

We used a hierarchical prior for center effects, and would like to determine if the variation between centers are greater than the variation due to the measurement error. Incorporating the insights from STA 721 project, we have

$$\beta_c \mid \sigma^2, \sigma_C^2, \lambda_c \sim N(0, \sigma^2\sigma_C^2/\lambda_c)$$

, where σ^2 is the within-group variance, and σ_C^2 is the between-group variance. We used a half-Cauchy prior on σ_L, and

$$\lambda_l \sim G(a/2, a/2)$$

. If a is small, then the posterior would be diffuse, and if a is large, then it would not allow for heavier tails than the normal for center effects, which would be a problem if we have any centers as "outliers". Thus, we need to set a a value which is not too large or too small, and a recommended choice for a is 9 as we learned from STA 721.

Since mixing is easily to be poor for sigma_L, then we removed all score variables, and introduced the observations which are initially removed because of the missing information of score variables. We proposed a model with no interactions, and no PCBs with a bunch of zeros. At the same time, we tried a model which

replaced the PCBs with the "total" PCB effect proxy as Youndsoo suggested, but we had a mixing problem for Sigma_L under this model.

```
data = Longnecker
data_age_na<- data[-which(is.na(data$pcb_028)), ]
```

```
full.lm<- lm(gestational_age ~ dde + pcb_028 + pcb_074 + pcb_105 + pcb_118 + pcb_153+ pcb_138 + pcb_180

Y = data_age_na$gestational_age
X0 = model.matrix(full.lm)
# remove X where the fitted value is NA
coef(full.lm)[is.na(coef(full.lm))]
```

```
## named numeric(0)
```

```
namesX0 = colnames(X0)

X1 = dplyr::select(data.frame(X0),starts_with("center"))
X2 = dplyr::select(data.frame(X0),-starts_with("center"))[,-1]
X = as.matrix(cbind(X1,X2))
model = function(){
  for (i in 1:n){
    Y[i] ~ dnorm(alpha + Xs[i,] %*% beta, phi)
  }

  alpha ~ dnorm(0, .000001*phi)
  phi ~ dgamma(.001, .001)
  sigma_L ~ dt(0,1,1)
  phi_L <- 1/((sigma_L^2)+.000001)
  tau ~ dgamma(.5, .5*n)
  sigma <- sqrt(1/phi)

  # beta for lab
  for (j in 1:p_center){
    lambda[j] ~ dgamma(a/2, a/2)
    beta[j] ~ dnorm(0, phi*phi_L*lambda[j]) # beta for lab starts from second column
  }

  # beta for others
  beta[(p_center+1):p] ~ dmnorm(rep(0,p-p_center), phi*tau*SSX2)
}
set.seed(1)
Xs = scale(X)
data = list(Y=Y, Xs=Xs, SSX2 = t(Xs[,(ncol(X1)+1):ncol(X)])%*%Xs[,(ncol(X1)+1):ncol(X)],
            n=length(Y), p_center=ncol(X1), p=ncol(X), a=9)
output = jags(data, inits=NULL,
              parameters.to.save=c("alpha","beta","sigma","sigma_L","lambda"),
              model=model, n.iter=10000)
```

```
## module glm loaded
```

```
## Compiling model graph
```

```
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 2379
##    Unobserved stochastic nodes: 27
##    Total graph size: 69248
##
## Initializing model
```

```
sim.matrix = output$BUGSoutput$sims.matrix


n = length(Y)
S = diag(c(apply(X,2,function(x) {var(x)})),nrow=ncol(X))
# Transform back original betas
beta.sim = sim.matrix[,2:26] %*% (solve(S)^0.5) #beta_s = S^0.5*beta
#max(abs(beta.sim %*% (S^0.5) - sim.matrix[,2:57])) # Checkings
# Transform back original alpha
alphaS.sim = sim.matrix[,1]
Xbar = apply(X,2,mean)
alpha.sim = alphaS.sim - beta.sim %*% Xbar
colnames(alpha.sim) = "(Intercept)"
colnames(beta.sim) = colnames(X)

quant5 = function(x) {round(quantile(x, c(0.025, 0.5, 0.975)),2)} ## 95% CI
apply(beta.sim, 2, quant5)
```

```
##       center10 center15 center31 center37 center45 center50 center55
## 2.5%     -0.08    -0.72    -0.24    -0.77    -0.03    -0.35    -0.27
## 50%       0.32    -0.21     0.26    -0.34     0.35     0.03     0.13
## 97.5%     0.85     0.15     0.87     0.02     0.85     0.45     0.63
##       center60 center66 center71 center82   dde pcb_028 pcb_074 pcb_105
## 2.5%     -0.35    -0.23    -0.46    -0.78 -0.02   -0.44   -1.43   -0.60
## 50%       0.07     0.06    -0.06    -0.29 -0.01    0.43   -0.53    1.31
## 97.5%     0.52     0.37     0.32     0.08  0.00    1.35    0.39    3.23
##       pcb_118 pcb_153 pcb_138 pcb_180 raceblack raceother triglycerides
## 2.5%    -1.00   -1.46   -1.07   -0.89     -1.08     -0.83         -0.01
## 50%     -0.27   -0.48    0.15    0.31     -0.75     -0.25          0.00
## 97.5%    0.46    0.55    1.37    1.61     -0.41      0.35          0.00
##       cholesterol maternal_age smoking_status
## 2.5%            0        -0.02          -0.32
## 50%             0         0.00          -0.09
## 97.5%           0         0.02           0.14
```

```
gelman.diag(as.mcmc(output))
```

```
## Potential scale reduction factors:
##
##            Point est. Upper C.I.
## alpha            1.00       1.00
## beta[1]          1.00       1.00
## beta[10]         1.00       1.00
## beta[11]         1.00       1.01
```

```
## beta[12]         1.00       1.00
## beta[13]         1.00       1.00
## beta[14]         1.00       1.00
## beta[15]         1.00       1.01
## beta[16]         1.00       1.00
## beta[17]         1.00       1.00
## beta[18]         1.00       1.00
## beta[19]         1.00       1.00
## beta[2]          1.00       1.00
## beta[20]         1.00       1.00
## beta[21]         1.00       1.00
## beta[22]         1.00       1.00
## beta[23]         1.00       1.01
## beta[24]         1.00       1.00
## beta[25]         1.00       1.00
## beta[3]          1.00       1.01
## beta[4]          1.00       1.00
## beta[5]          1.00       1.01
## beta[6]          1.00       1.00
## beta[7]          1.00       1.00
## beta[8]          1.00       1.00
## beta[9]          1.00       1.00
## deviance         1.00       1.01
## lambda[1]        1.00       1.00
## lambda[10]       1.00       1.00
## lambda[11]       1.00       1.00
## lambda[2]        1.00       1.00
## lambda[3]        1.00       1.00
## lambda[4]        1.00       1.00
## lambda[5]        1.00       1.00
## lambda[6]        1.00       1.00
## lambda[7]        1.00       1.00
## lambda[8]        1.00       1.00
## lambda[9]        1.01       1.02
## sigma            1.00       1.00
## sigma_L          2.29       4.50
##
## Multivariate psrf
##
## 1.76
```

```
### mixing is pretty good
```

```
coefs_for_table<- apply(beta.sim, 2, quant5)
knitr::kable(t(coefs_for_table), format = "latex")
```

	2.5%	50%	97.5%
center10	-0.08	0.32	0.85
center15	-0.72	-0.21	0.15
center31	-0.24	0.26	0.87
center37	-0.77	-0.34	0.02
center45	-0.03	0.35	0.85
center50	-0.35	0.03	0.45
center55	-0.27	0.13	0.63
center60	-0.35	0.07	0.52
center66	-0.23	0.06	0.37
center71	-0.46	-0.06	0.32
center82	-0.78	-0.29	0.08
dde	-0.02	-0.01	0.00
pcb_028	-0.44	0.43	1.35
pcb_074	-1.43	-0.53	0.39
pcb_105	-0.60	1.31	3.23
pcb_118	-1.00	-0.27	0.46
pcb_153	-1.46	-0.48	0.55
pcb_138	-1.07	0.15	1.37
pcb_180	-0.89	0.31	1.61
raceblack	-1.08	-0.75	-0.41
raceother	-0.83	-0.25	0.35
triglycerides	-0.01	0.00	0.00
cholesterol	0.00	0.00	0.00
maternal_age	-0.02	0.00	0.02
smoking_status	-0.32	-0.09	0.14

```
beta.sim_center<- dplyr::select(data.frame(beta.sim),starts_with("center"))

center_coef<- matrix(nrow=3, ncol=11)
for (i in 1:3){
  center_coef[i,]<- apply(alpha.sim, 2, quant5)[i] + apply(beta.sim_center, 2, quant5)[i,]
}
colnames(center_coef)<- colnames(beta.sim_center)
rownames(center_coef)<- c("2.5%", "50%", "97.5")
center_coef<- as.data.frame(center_coef)
center_coef$center5<- apply(alpha.sim, 2, quant5)
coefs_for_table<- t(center_coef)
coefs_for_table
```

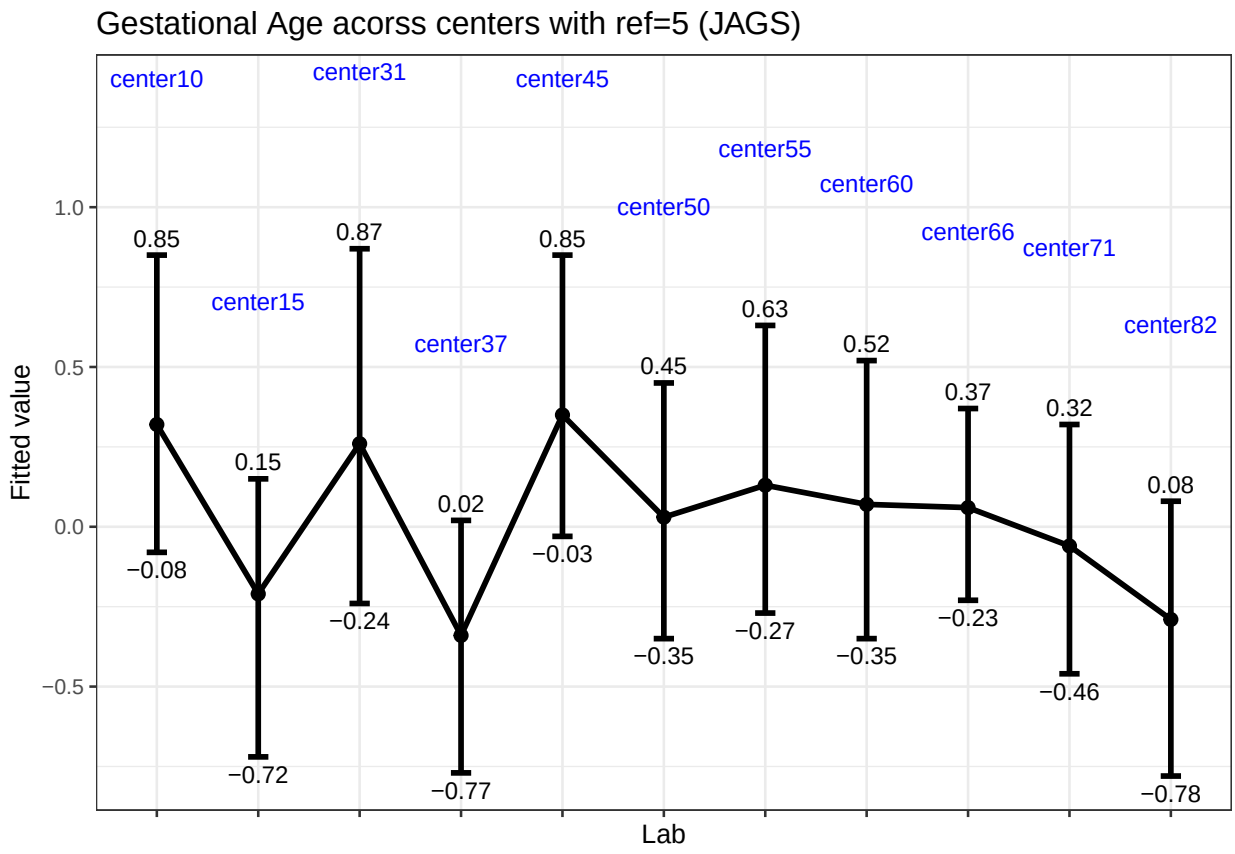
```
##           2.5%   50%  97.5
## center10 39.72 40.80 42.05
## center15 39.08 40.27 41.35
## center31 39.56 40.74 42.07
## center37 39.03 40.14 41.22
## center45 39.77 40.83 42.05
## center50 39.45 40.51 41.65
## center55 39.53 40.61 41.83
## center60 39.45 40.55 41.72
## center66 39.57 40.54 41.57
## center71 39.34 40.42 41.52
## center82 39.02 40.19 41.28
## center5  39.80 40.48 41.20
```

```
knitr::kable(coefs_for_table, format = "latex")
```

	2.5%	50%	97.5
center10	39.72	40.80	42.05
center15	39.08	40.27	41.35
center31	39.56	40.74	42.07
center37	39.03	40.14	41.22
center45	39.77	40.83	42.05
center50	39.45	40.51	41.65
center55	39.53	40.61	41.83
center60	39.45	40.55	41.72
center66	39.57	40.54	41.57
center71	39.34	40.42	41.52
center82	39.02	40.19	41.28
center5	39.80	40.48	41.20

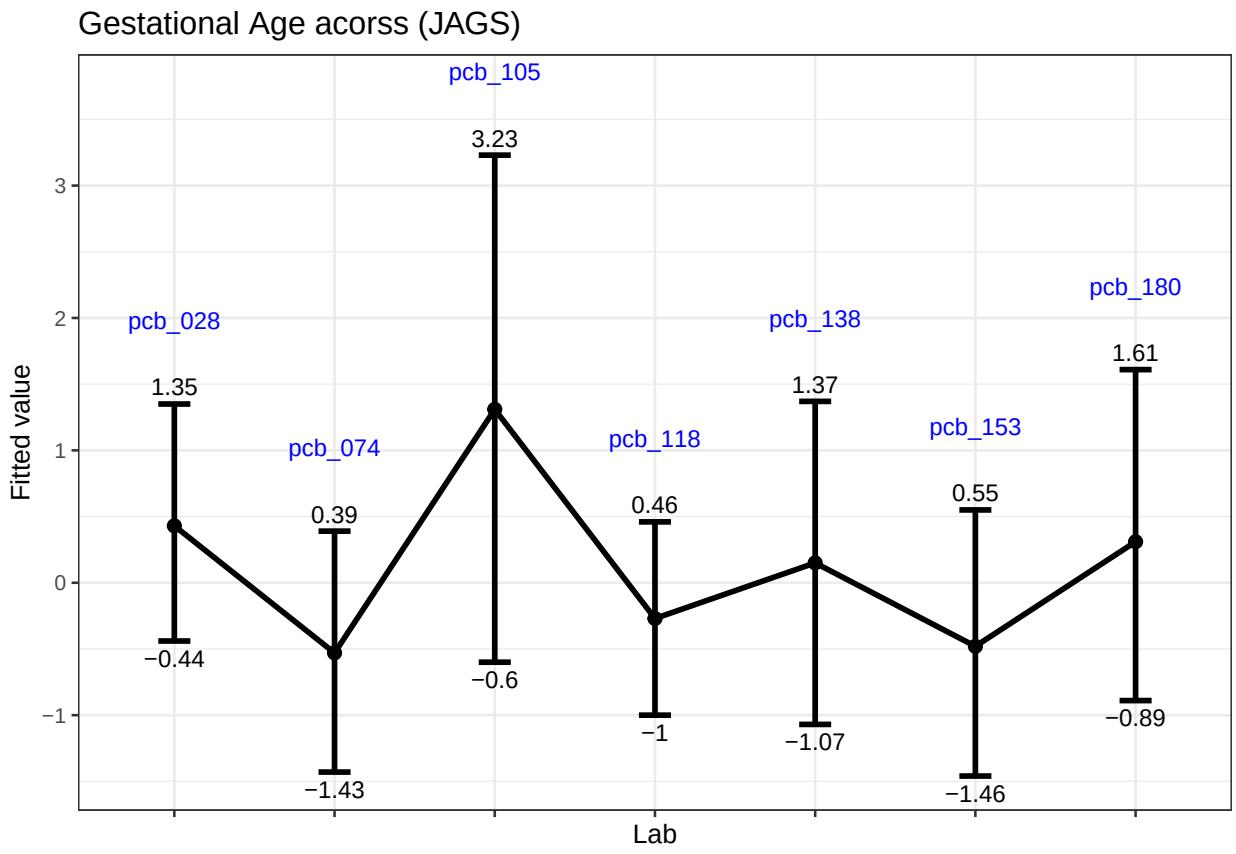
```
center = dplyr::select(data.frame(beta.sim),starts_with("center"))
center_fit<- cbind(lwr= apply(center, 2, quant5)[1,], fit= apply(center, 2, quant5)[2,], upr=apply(cent

ggplot(as.data.frame(center_fit), aes(rownames(center_fit),
  fit, size=10, group=1, ylim=max(upr)+0.8)) +
  theme_bw(base_size=10)+
  geom_point(size=2)+
  geom_line(size=1)+
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.2, size=1)+
  xlab("Lab")+
  ylab("Fitted value")+
  ggtitle("Gestational Age acorss centers with ref=5 (JAGS)")+
  theme(axis.text.x=element_blank())+
  geom_text(aes(label = round(lwr,2), y = lwr), vjust = 1.5, size=3) +
  geom_text(aes(label = round(upr,2), y = upr), vjust = -0.5, size=3) +
  geom_text(aes(label = rownames(center_fit), y = upr+0.5), vjust = -0.5, size=3, col="blue")
```




```
pcb = dplyr::select(data.frame(beta.sim),starts_with("pcb_"))
pcb_fit<- cbind(lwr= apply(pcb, 2, quant5)[1,], fit= apply(pcb, 2, quant5)[2,], upr=apply(pcb, 2, quant5

ggplot(as.data.frame(pcb_fit), aes(rownames(pcb_fit),
  fit, size=10, group=1, ylim=max(upr)+0.8)) +
  theme_bw(base_size=10)+
  geom_point(size=2)+
  geom_line(size=1)+
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.2, size=1)+
  xlab("Lab")+
  ylab("Fitted value")+
  ggtitle("Gestational Age acorss (JAGS)")+
  theme(axis.text.x=element_blank())+
  geom_text(aes(label = round(lwr,2), y = lwr), vjust = 1.5, size=3) +
  geom_text(aes(label = round(upr,2), y = upr), vjust = -0.5, size=3) +
  geom_text(aes(label = rownames(pcb_fit), y = upr+0.5), vjust = -0.5, size=3, col="blue")
```




```
### center 10 31 45 higher than center 5; center 15 37 82 lower than center 5
### pcbs all cover 0, but 105, 028 higher, and 074 lower
```