Case Study #1

Shrey Gupta, Frances Hung, Ezinne Nwankwo

Abstract

1 Introduction

We are interested in studying how chemical exposure, in particular exposure to Dichlorodiphenyldichloroethylene (DDE) and Polychlorinated Biphenyls (PCBs), relates to risk of pre-mature delivery. Early delivery is known to be associated with an increased risk of adverse outcomes for the child, such as risk of morbidity and mortality.

DDE and PCBs are commonly used to treat crops and protect them from predation. As a result, they are prevalent in the environment and can build up in fatty deposits in human tissues, thus leading to negative health outcomes for humans. We hope to provide some measure of how these chemicals are related to preterm delivery, while controlling for other covariates that may confound the exposure-outcome relationship.

1.1 The Data

For this study, we use a subsample of 2,380 women and children from Longnecker, et al. (2001) that was initially collected by the National Collaborative Perinatal Project. The data include various demographic variables (race, age, and socioeconomic markers), smoking status, concentration doses of DDE and PCBs due to exposure, and cholesterol and triglycerides levels. Given that the outcome variable is not normally distributed and has a heavy right tail since most women carry to full term (>= 37 weeks), we decided to define a binary outcome variable for pre-term and not pre-term with a cut-off of 36 weeks or fewer. Due to 92% missingness, we also decided to completely omit the variable measuring albumin from our analysis. We also omitted the three variables related to socioeconomic status and education due to their lack of interpretability even after imputation. Getting rid of these covariates in our analysis removed most of the missingness from our dataset. Lastly, we noticed that the PCBs variables were highly correlated with each other so instead of including all 11 variables into the model, we considered the average, minimum, and maximum PCB exposures for each patient.

2 Materials & Methods

Since linear model assumptions (namely, normality of residuals) were not satisfied in this dataset, we instead chose to implement a logistic model. To satisfy the assumptions needed for logistic models, we modified our data. The model predicts whether an observation is pre-term (<=36 weeks) or around normal (>36 weeks), so the dependent variable, gestational age, is changed to be binary. Our observations are assumed to be independent from one another, and we use variation inflation factors and Bayesian Model Averaging (described later) to get rid of multicollinearity. One assumption, that the predictors have a linear relationship with the logit function, was not totally satisfied, but our model still managed to capture inferential trends; we are looking for a model which captures the general relationship between DDE and gestational age, not an accurate predictive model.

We first used Bayesian Model Averaging for generalized linear models to explore variable importance. Key variables with significant probabilities of inclusion were triglycerides, centers, and DDE, and the noninclusion of other variables like maternal age and smoking status were corroborated by running a full naive GLM model. We double check the multicollinearity of our chosen variables by looking at variable inflation factors and conclude that these variables are not significantly correlated. From our EDA analysis showing differences in gestational ages but similar racial trends across centers, we decided to add a random-effect intercept to the logistic model based on centers. Because the goal of this analysis was to assess effects of DDE and PCB on gestational age, we also included the average of the PCB variates as a covariate in our model. Our final model that we implemented was a generalized linear mixed effects model with logit link function and a random intercept across centers:

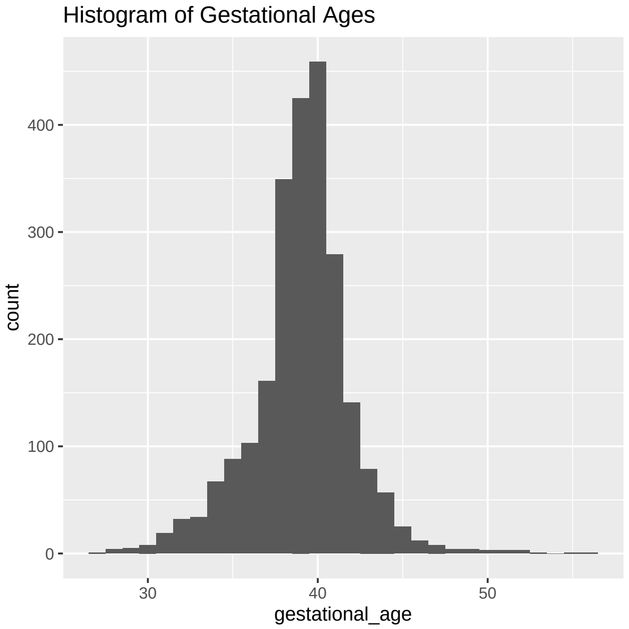
$$Pr(Y_{ij} = 1) = X_{ij}\beta + \alpha_j^{center}$$
$$\alpha_j \sim N(0, \sigma^2)$$

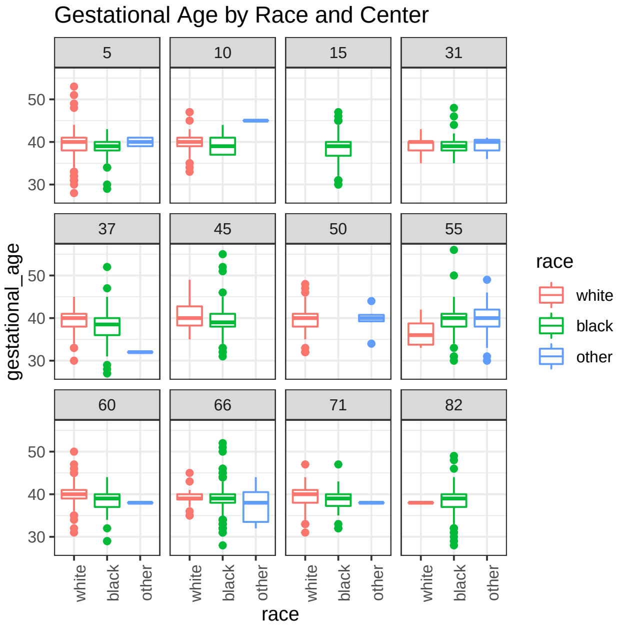
where Y_{ij} is the binary outcome variable for patient i in hospital j, X_{ij} is the model matrix where the first column is a column of 1's for the fixed intercept and the remaining columns are the values for the covariates included in our model (i,e. race, dde, average pcb, and triglycerides).

3 Results

3.1 Exploratory Data Analysis

From the histogram below on the left, the dependent variable does not seem to follow a normal distribution. We investigate this futher with a normal QQ plot (See Appendix) and determine that a linear regression model is not suitable since the residuals are not normally distributed. We also look at the relationship between predictors and the outcome variable. The most impactful finding is shown in the plot on the right which visualizes distribution of gestational age across centers and races. We see that there is heterogeneity across centers that we should account for in our model.





3.2 Main Results

The exploratory analysis revealed that there is significant heterogeneity across centers which immediately made us consider a multilevel model to account for such variation and improve estimates of other covariates. Additionally, given that some centers did not even treat certain races of mothers (i.e. centers 15, 45, 82), if we just included centers as predictors in our model then those estimates would not necessarily be trustworthy. We know there are differences across centers and we want to control for that. We also noticed some slight differences in race distributions and therefore considered a random slope model as well but it turned out not be the best in terms of BIC.

Table 1 describes our final model estimates and uncertainty quantification about those estimates.

```
## boundary (singular) fit: see ?isSingular
```

	Estimates	Lower 2.5% CI	Upper 97.5% CI
(Intercept)	3.24	2.64	3.85
I(triglycerides/100)	-0.17	-0.33	0.01
I(dde/100)	-0.72	-1.34	-0.07
avg_pcb	-0.52	-1.48	0.48
raceblack	-0.48	-0.93	-0.01
raceother	-0.81	-1.49	-0.09

We also wanted to emphasize that we end up with different estimates for the predictors when we include random intercepts for centers versus just including a categorical predictor for center in a logistic model. We notice that the estimate for DDE increases in absolute value and remains significant, the estimate for Avg PCB decreases in absolute value and remains not significant, and the race indicators become significant in the random intercept model. We believe that this is further justification for our model because when we do not account for center heterogeneity, the effect of some of our predictors become obscured.

3.3 Model Selection & Diagnostics

We focused on four models and conducted model selection scheme using BIC to determine the best fit. We found that the BIC is lowest for the random intercept model only model.

Model	BIC
Random Intercept	1631.35
Random Slope	1665.54
Random Slope + Random Int	1673.31
No Random	1676.19

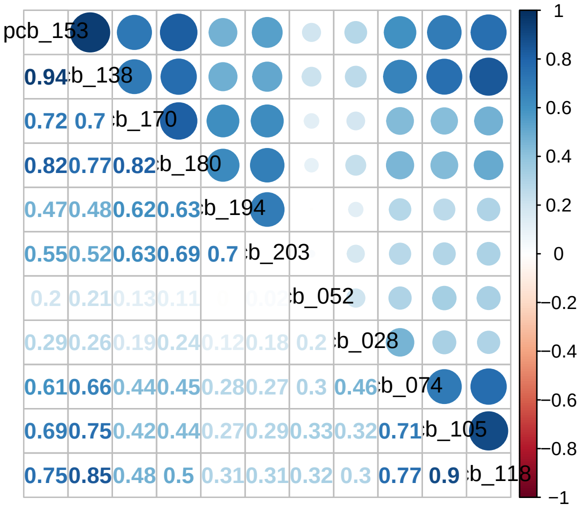
4 Discussion

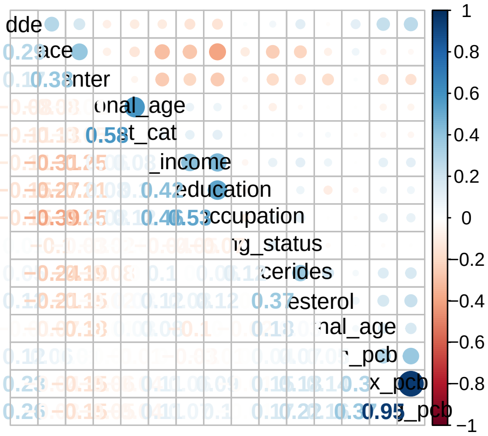
Our results find that the effect for average PCB is not significant, meaning there is no evidence of an association between PCB exposure and risk of pre-term delivery. However, higher DDE exposure is associated with higher risk of pre-term delivery. A one unit increase in DDE is associated with decreasing the expected odds of having a full-term pregnancy by approximately a factor of 2 (0.72 decrease in the expected log odds), holding everything else fixed. In addition, we find several other interesting pieces of insight. Higher triglycerides are associated with a higher risk of pre-term delivery, as are being a non-white mother.

There are various advantages and disadvantages of the approach we took. On one hand, regression is highly interpretable, and interpretability is important for disciplines like the health and sciences. (We also used the min, max, and average PCB exposures instead of doing PCA since the former is more interpretable.) Furthermore, using a random intercept model allows us to take into account the heterogeneity across centers in our model. On the flip side, as discussed in the results, not all of the assumptions for logistic regression were satisfied in this study (particularly the linearity assumption). Logistic regression also only gives a binary outcome: pre-term or full-term, which may not be as useful as the outcome predicted from ordinal, quantile, or linear regression, which would provide more specificity on the time range in which delivery occurs.

These results are consistent with some of the trends we saw in our exploratory plots and with current literature surrounding pre-term deliveries. Future directions for analysis include (1) sensitivity analysis on the number of weeks that defines a pre-term birth, (2) multiple category outcome modeling using Bayesian GLMM, and (3) accounting for natural ordering in outcome via a proportional odds model.

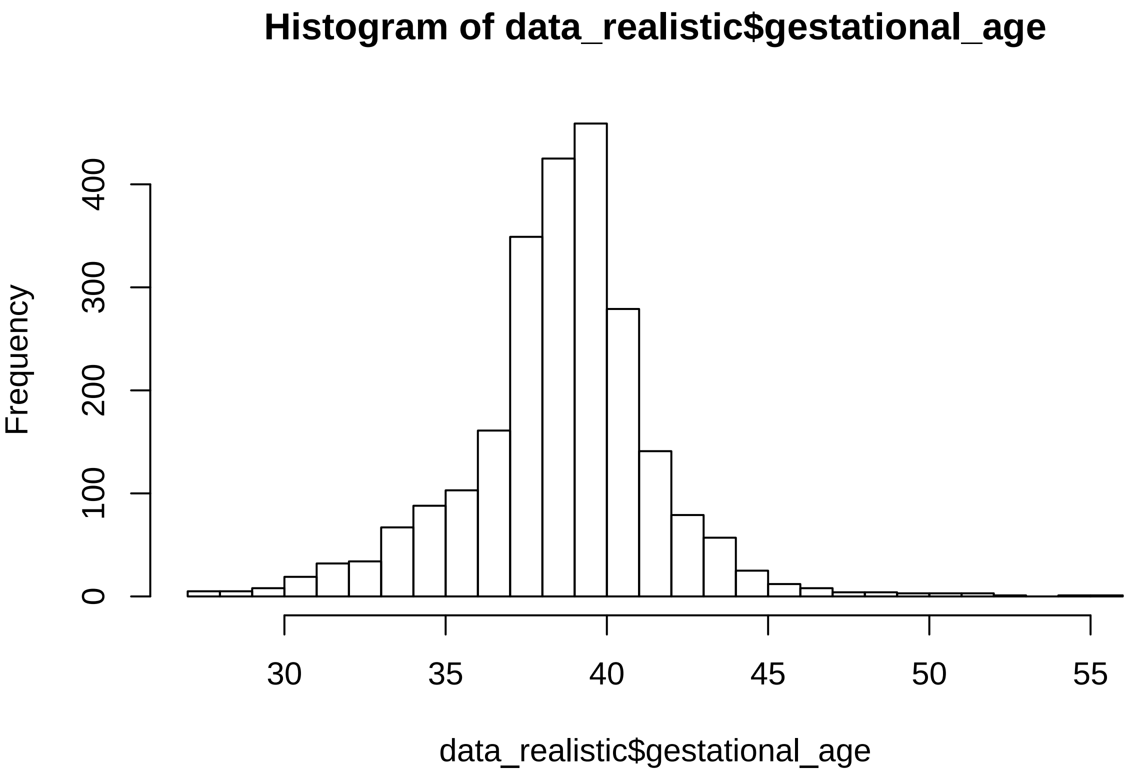
Appendix: Figures & Analysis





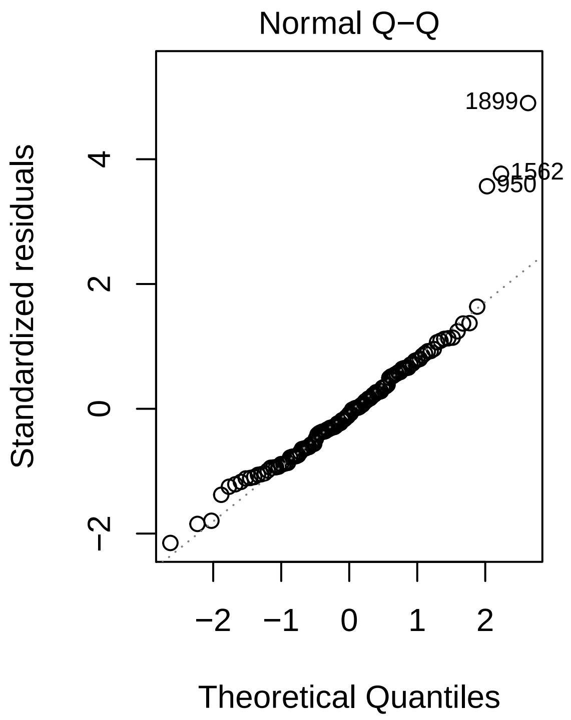
We observe that PCB variations are positively correlated with one another, and that certain groups of variables are also correlated (education, occupation, and income; triglycerides and cholesterol; race and center; race and DDE; maternal age and triglycerides, etc.).

```
data_realistic <- data %>% filter(gestational_age <= 60)
hist(data_realistic$gestational_age, breaks=25)
```



```
lm_model <- lm(gestational_age ~ ., data_realistic)
par(mfrow=c(1,2))
plot(lm_model, which=2)
```

```
## Warning: not plotting observations with leverage one:
##   104, 105
```



```
# summary stats for outcome by center
samp_stats <- data %>% group_by(center) %>%
  summarise(nj=n(), avg_gest=mean(gestational_age),
            variance=var(gestational_age)) %>% data.frame()
```

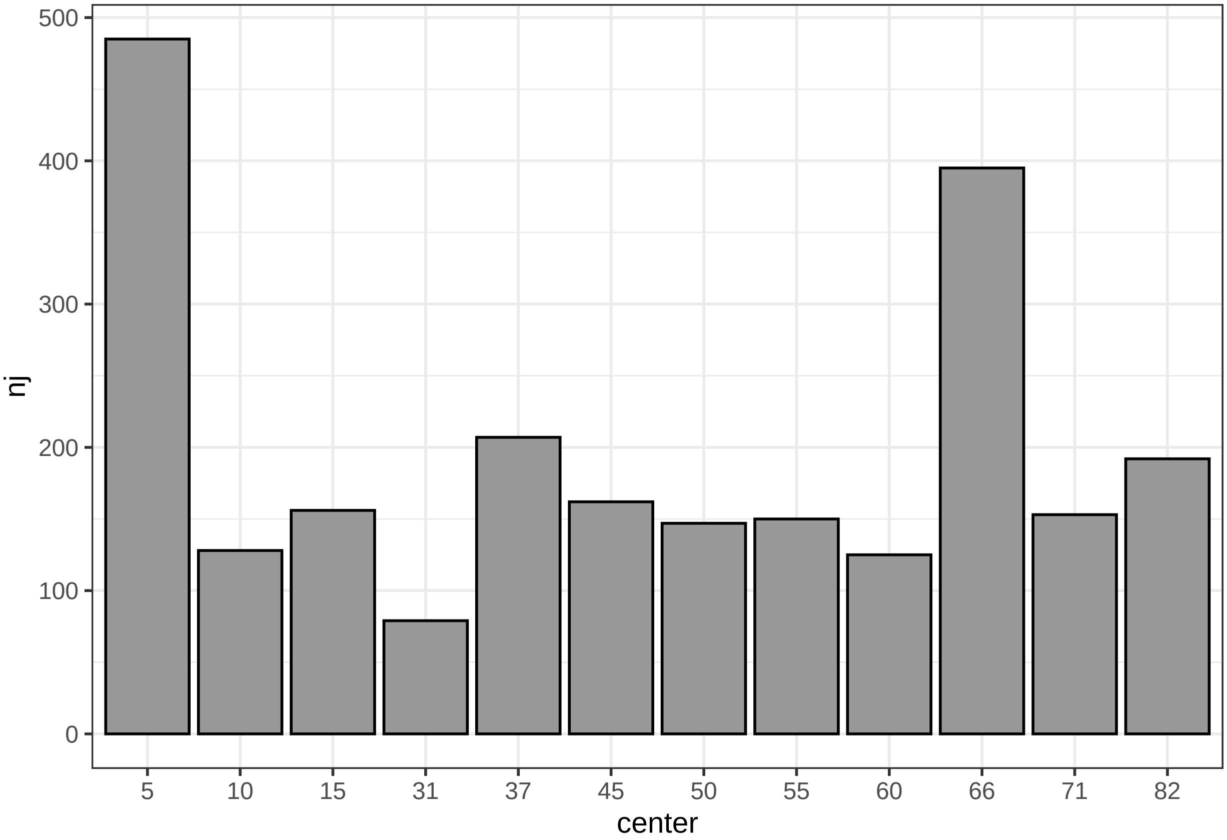
```
## Warning: Grouping rowwise data frame strips rowwise nature
```

```
data %>% group_by(race) %>%
  summarise(nj=n(), avg_gest=mean(gestational_age),
            variance=var(gestational_age)) %>% data.frame()
```

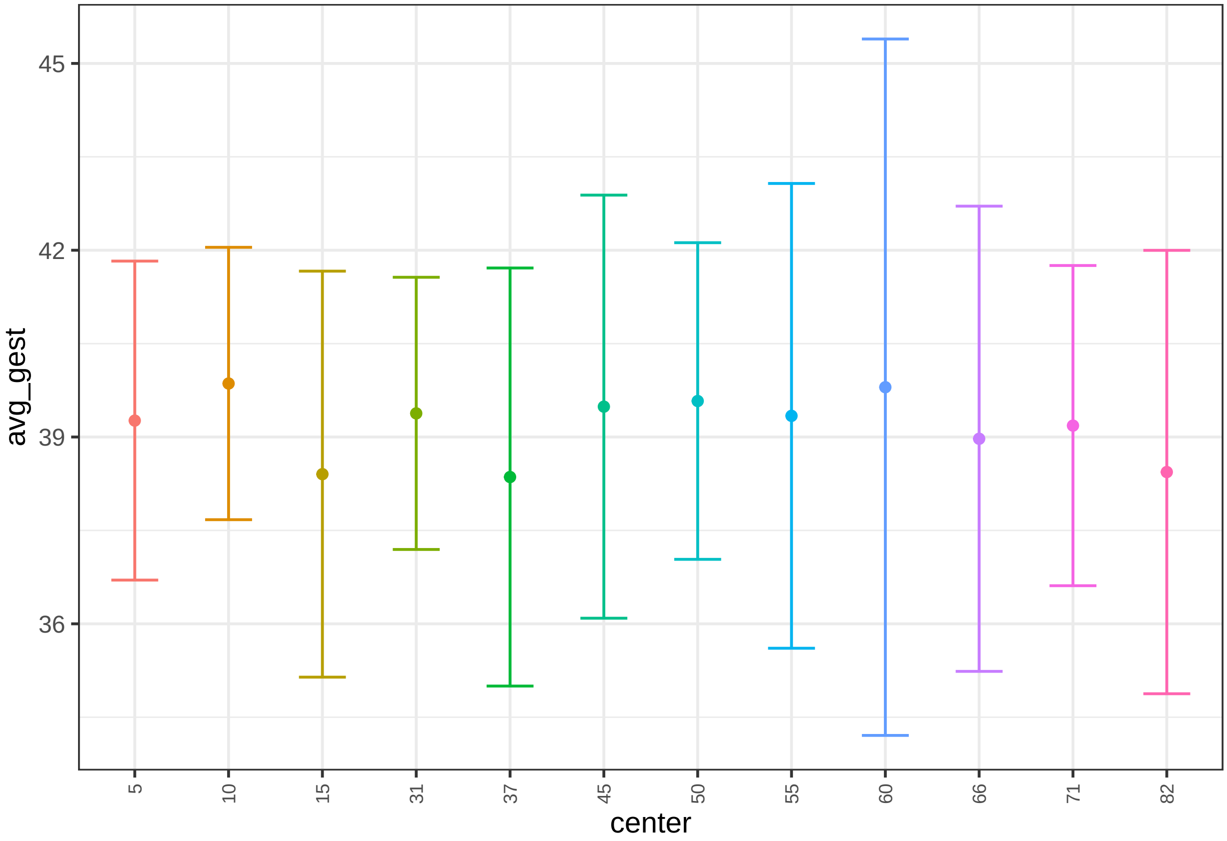
```
## Warning: Grouping rowwise data frame strips rowwise nature
```

```
##    race   nj avg_gest  variance
## 1 white 1032 39.45833  6.723771
## 2 black 1223 38.76043 12.388547
## 3 other  124 39.69355 31.872804
```

```
samp_stats %>% ggplot(aes(x=center, y=nj)) +
  geom_bar(stat="identity", color="black", fill="#999999") +
  theme_bw()
```



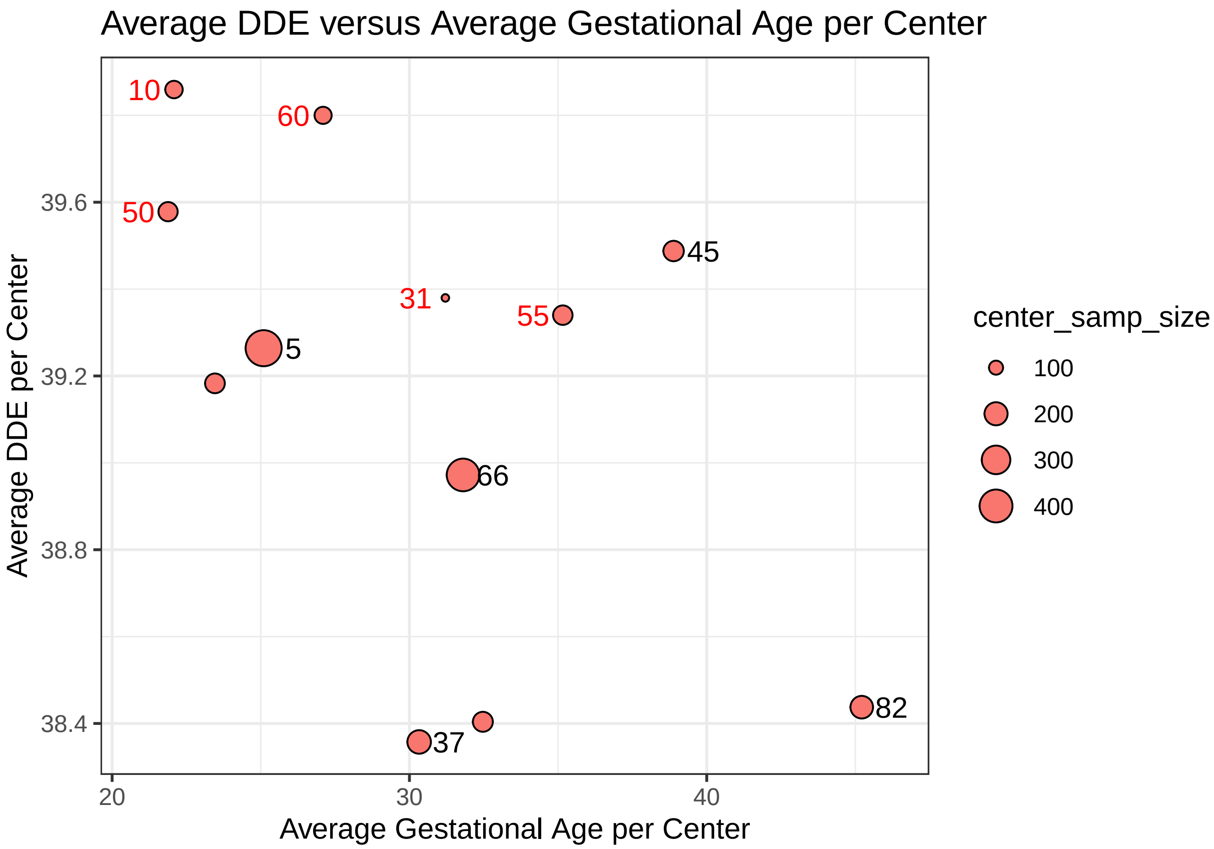
```
# plot summary stats
g1 <- samp_stats %>% mutate(se=sqrt(variance)) %>%
  ggplot(aes(x=center, y=avg_gest, color=center)) +
  geom_point() +
  theme_bw() +
  geom_errorbar(aes(ymin=avg_gest - se, ymax=avg_gest + se), width=.5) +
  theme(legend.position="none",
        axis.text.x=element_text(angle=90, size=7, vjust=0.5, hjust=1))
g1
```



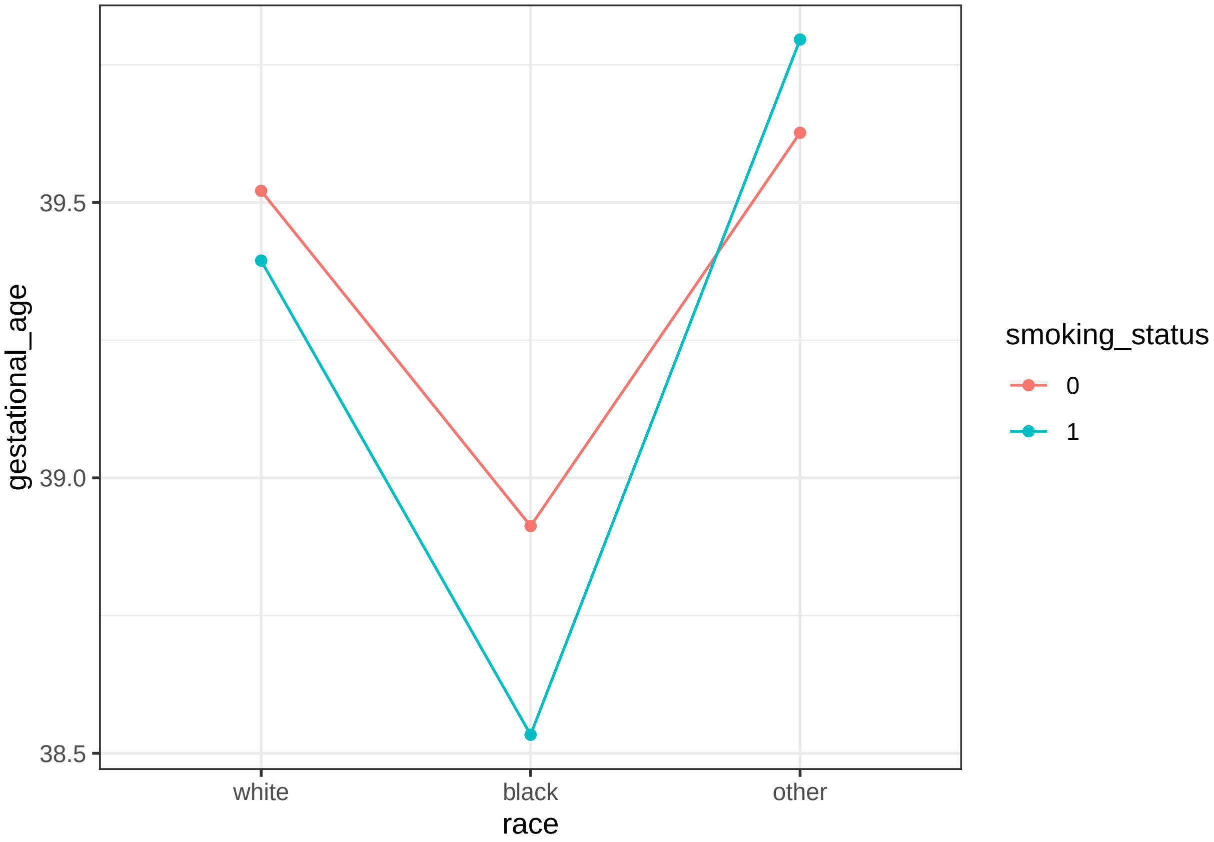
```
# summary stats for outcome and dde by center
samp_stats <- data %>% group_by(center) %>%
  summarise(center_samp_size=n(), avg_gest=mean(gestational_age),
            avg_dde=mean(dde))
```

```
## Warning: Grouping rowwise data frame strips rowwise nature
```

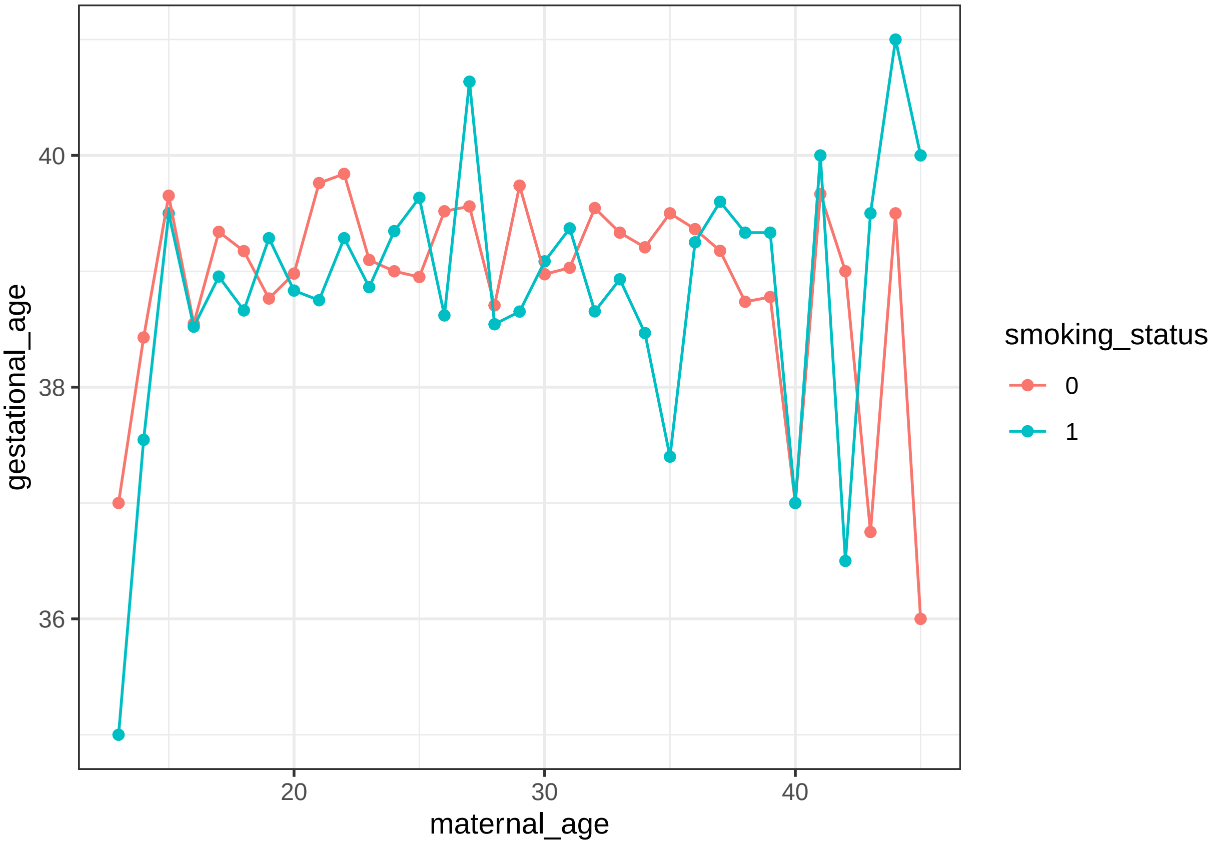
```
top_5 <- samp_stats %>% top_n(n=5, center_samp_size)
bot_5 <- samp_stats %>% top_n(n=-5, center_samp_size)

samp_stats %>%
  ggplot(aes(x=avg_dde, y=avg_gest, size=center_samp_size)) +
  geom_point(shape=21, fill="#F8766D") +
  annotate("text", x=top_5$avg_dde + 1, y=top_5$avg_gest, label=top_5$center) +
  annotate("text", x=bot_5$avg_dde - 1, y=bot_5$avg_gest, label=bot_5$center,
           colour="red") +
  theme_bw() +
  labs(y="Average DDE per Center", x="Average Gestational Age per Center",
       title="Average DDE versus Average Gestational Age per Center")
```



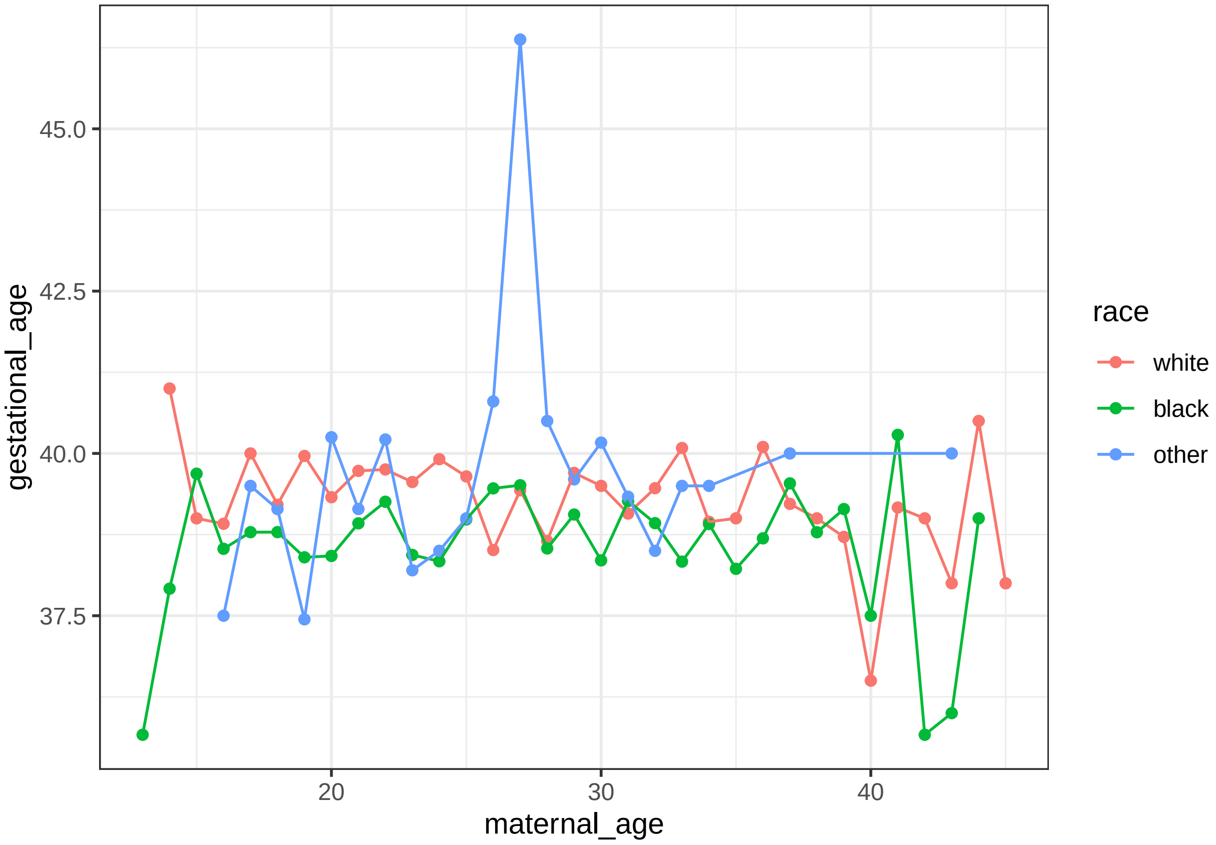
```r
# relationship between race and smoking status
# no constant (or parallel), so indicates interactive effect
data %>% ggplot() +
  aes(x=race, color=smoking_status, group=smoking_status, y=gestational_age) +
  stat_summary(fun.y=mean, geom="point") +
  stat_summary(fun.y=mean, geom="line") +
  theme_bw()
```



```
# relationship between age and smoking status
# no constant (or parallel), so indicates interactive effect
data %>% ggplot() +
  aes(x=maternal_age, color=smoking_status, group=smoking_status, y=gestational_age) +
  stat_summary(fun.y=mean, geom="point") +
  stat_summary(fun.y=mean, geom="line") +
  theme_bw()
```



```
# relationship between race and age
# no constant (or parallel), so indicates interactive effect
data %>% ggplot() +
  aes(x=maternal_age, color=race, group=race, y=gestational_age) +
  stat_summary(fun.y=mean, geom="point") +
  stat_summary(fun.y=mean, geom="line") +
  theme_bw()
```



BMA for GLM

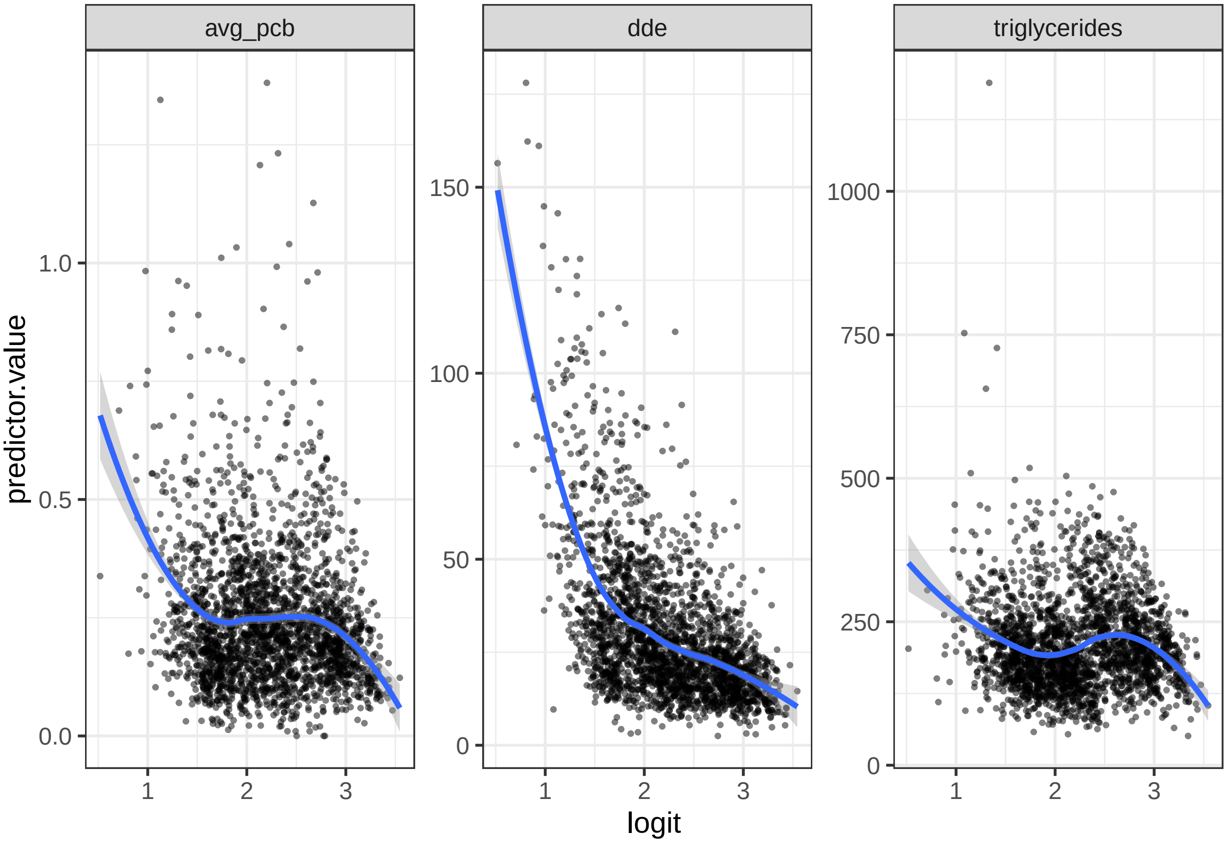
```
# remove uninterpretable variables and albumin
data_realistic <- data_realistic %>%
  select(-albumin, -score_education, -score_income, -score_occupation) %>% drop_na()

data_realistic <- data_realistic %>%
  mutate(pcb=data_realistic %>% select(pcb=starts_with("pcb")) %>% rowSums()) %>%
  select(-starts_with("pcb_"))

data_realistic <- data_realistic %>% mutate(gestational_age=case_when(gestational_age <= 36 ~ 0,
                                                                      gestational_age > 36 ~ 1))

bic.glm(x=data_realistic %>% select(-gestational_age), y=data_realistic$gestational_age,
        glm.family="binomial")
```

```
##
## Call:
## bic.glm.data.frame(x = data_realistic %>% select(-gestational_age),     y = data_realistic$gestationa
##
##
##  Posterior probabilities(%):
##            dde  triglycerides           race    maternal_age smoking_status
##          100.0           46.7            0.0             0.0           39.7
##    cholesterol         center       gest_cat         min_pcb        max_pcb
##            0.0            0.0          100.0            0.0            0.0
##        avg_pcb            pcb
##            0.0            0.0
##
##  Coefficient posterior expected values:
##         (Intercept)                  dde        triglycerides
##          -17.789217            -0.016709            -0.001392
##           raceblack            raceother         maternal_age
##            0.000000             0.000000             0.000000
##     smoking_status1          cholesterol             center10
##           -0.215587             0.000000             0.000000
##            center15             center31             center37
##            0.000000             0.000000             0.000000
##            center45             center50             center55
##            0.000000             0.000000             0.000000
##            center60             center66             center71
##            0.000000             0.000000             0.000000
##            center82  gest_catnot_preterm              min_pcb
##            0.000000            21.725299             0.000000
##             max_pcb              avg_pcb                  pcb
##            0.000000             0.000000             0.000000
```

##Checking Logistic Model Assumptions

Linearity of Covariates to Logit:

```r
ggplot(mydata, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```



Multicollinearity doesn't appear to be a concern.

```
car::vif(m1)
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4

##                           GVIF Df GVIF^(1/(2*Df))
## I(triglycerides/100) 1.082911  1        1.040630
## I(dde/100)           1.162908  1        1.078382
## avg_pcb              1.189217  1        1.090512
## race                 1.128089  2        1.030590
```