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 in data & remove data point with missing PCB information
dat <- readRDS("Longnecker.rds")
which(is.na(dat$pcb_028)==TRUE)
```

```
## [1] 1861
```

```
dat[1861,]
```

```
##        dde pcb_028 pcb_052 pcb_074 pcb_105 pcb_118 pcb_153 pcb_170 pcb_138
## 1861 16.62      NA      NA      NA      NA      NA      NA      NA      NA
##      pcb_180 pcb_194 pcb_203 albumin triglycerides  race score_education
## 1861      NA      NA      NA      NA           145 black              22
##      score_income score_occupation maternal_age smoking_status cholesterol
## 1861           15                5           14              1         213
##      gestational_age center
## 1861              42     37
```

```
dat <- dat[-1861,]
dat$race <- as.factor(dat$race)
dat$smoking_status <- as.factor(dat$smoking_status)
dat$center <- as.factor(dat$center)
attach(dat)
```

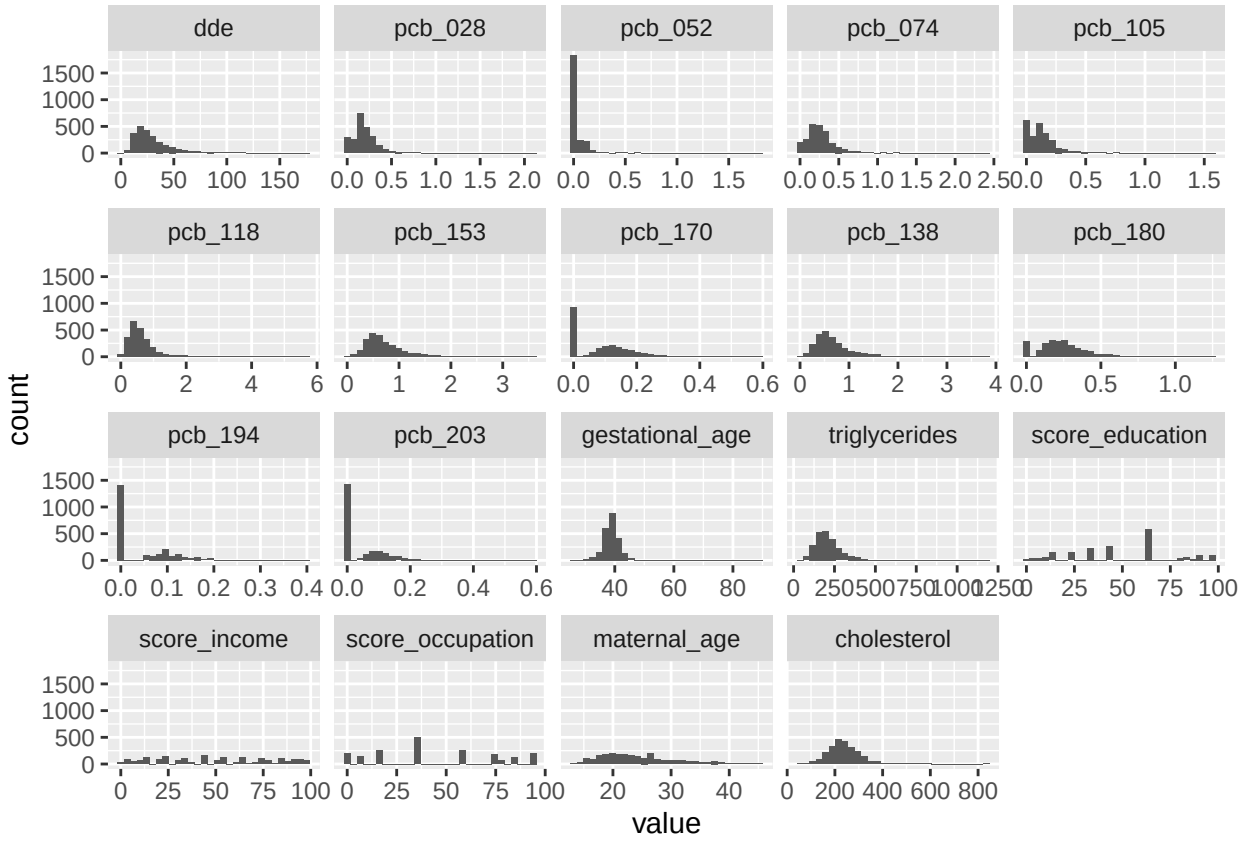
```
# Histograms
library(reshape2)
library(ggplot2)
d <- melt(dat[,c(1:12,22,14,16,17,18,19,21)])
```

```
## No id variables; using all as measure variables
```

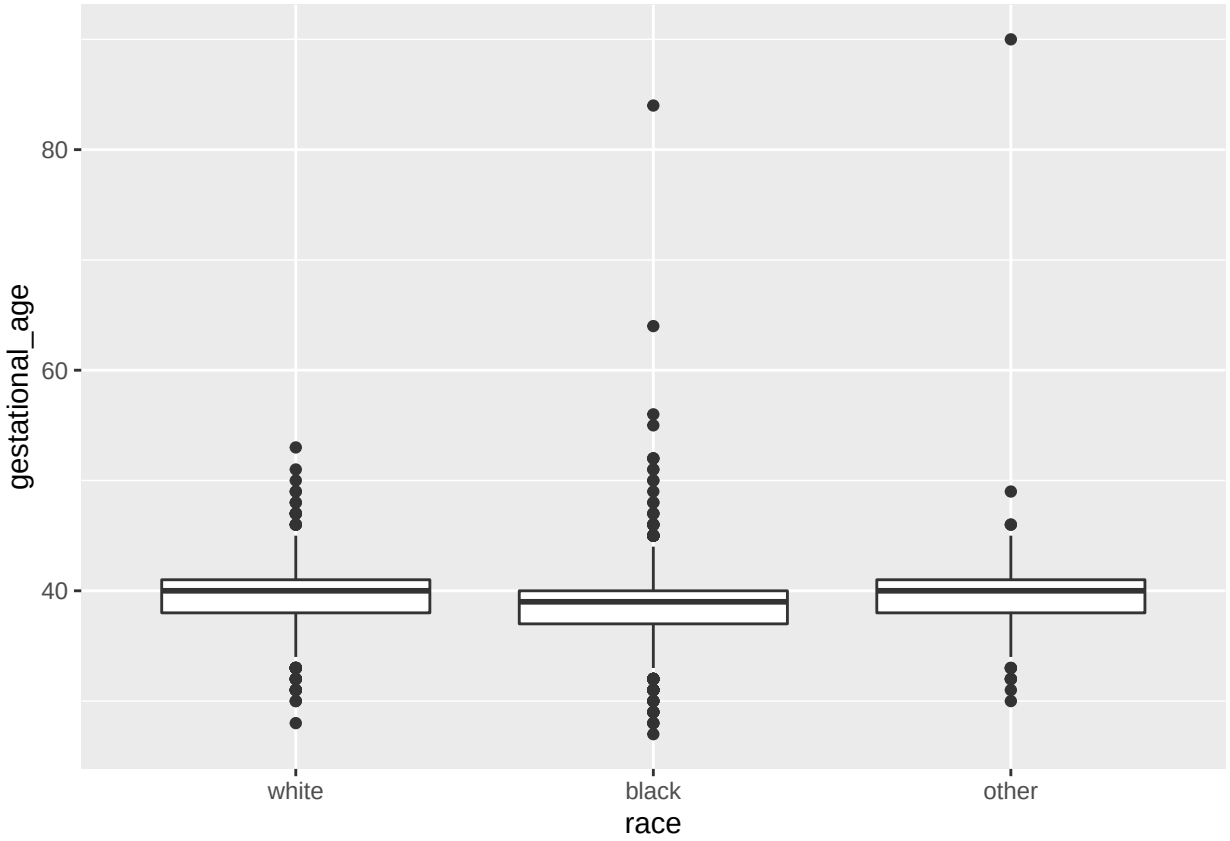
```
ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

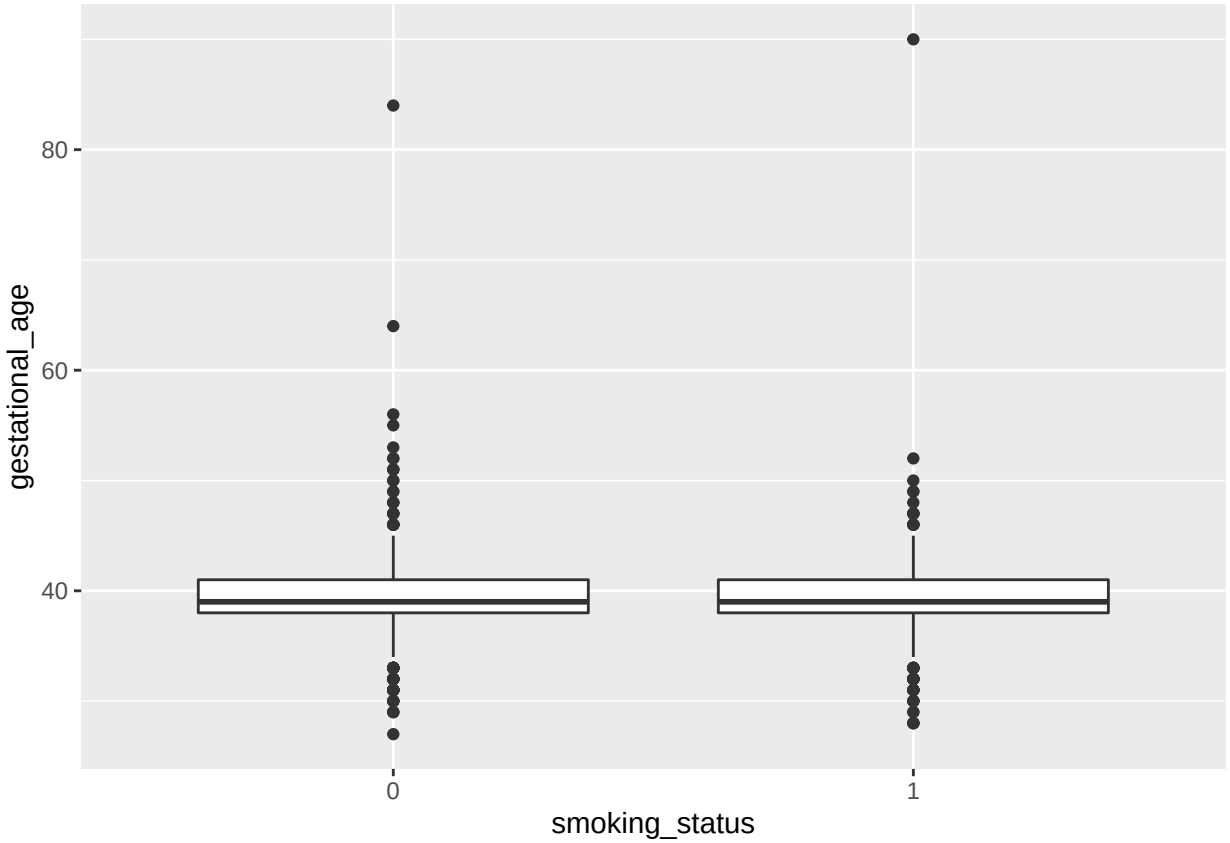
```
## Warning: Removed 1475 rows containing non-finite values (stat_bin).
```



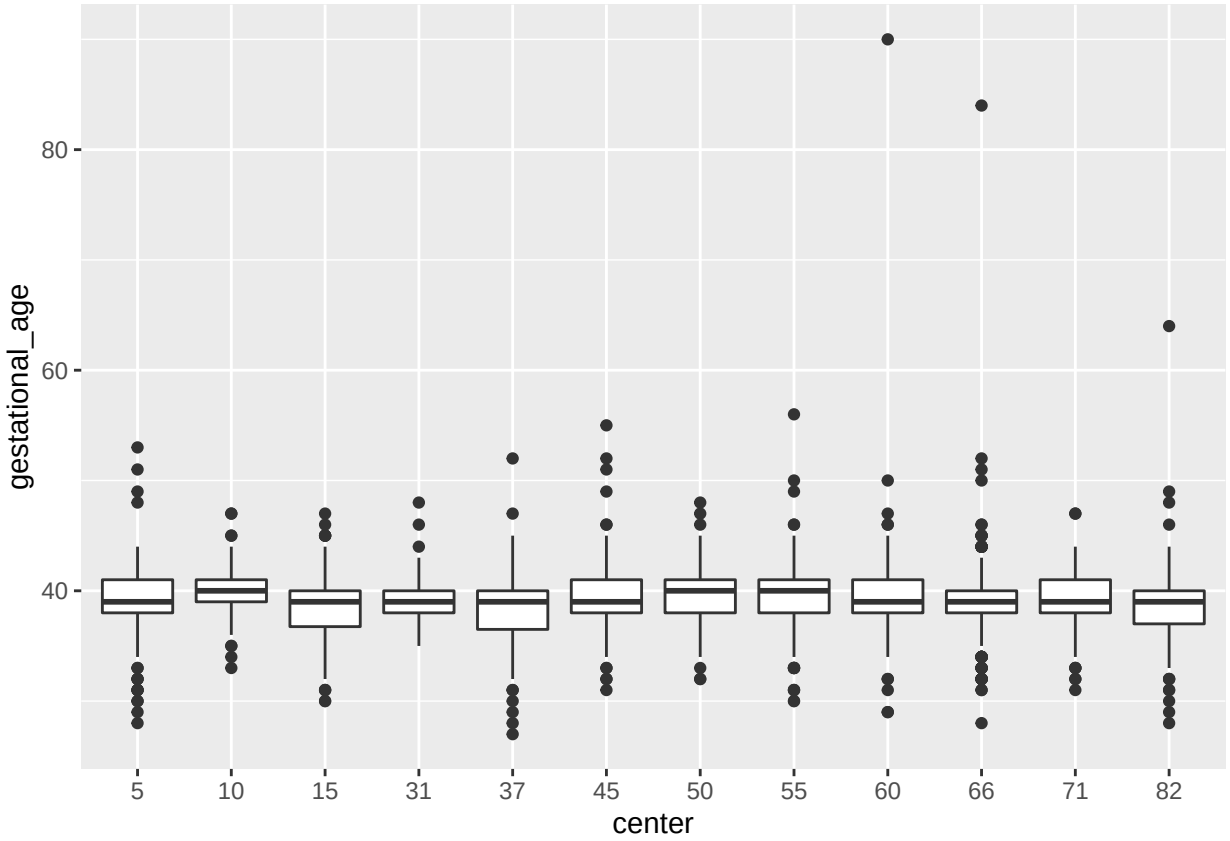

```
# # Log-transformation
# library(dplyr)
# log_d <- d
# log_d <- mutate(log_d,value = log(value))
# ggplot(log_d,aes(x = value)) +
#     facet_wrap(~variable,scales = "free_x") +
#     geom_histogram()

# Boxplots
ggplot(dat, aes(group=race, x=race, y=gestational_age)) +
  geom_boxplot()
```



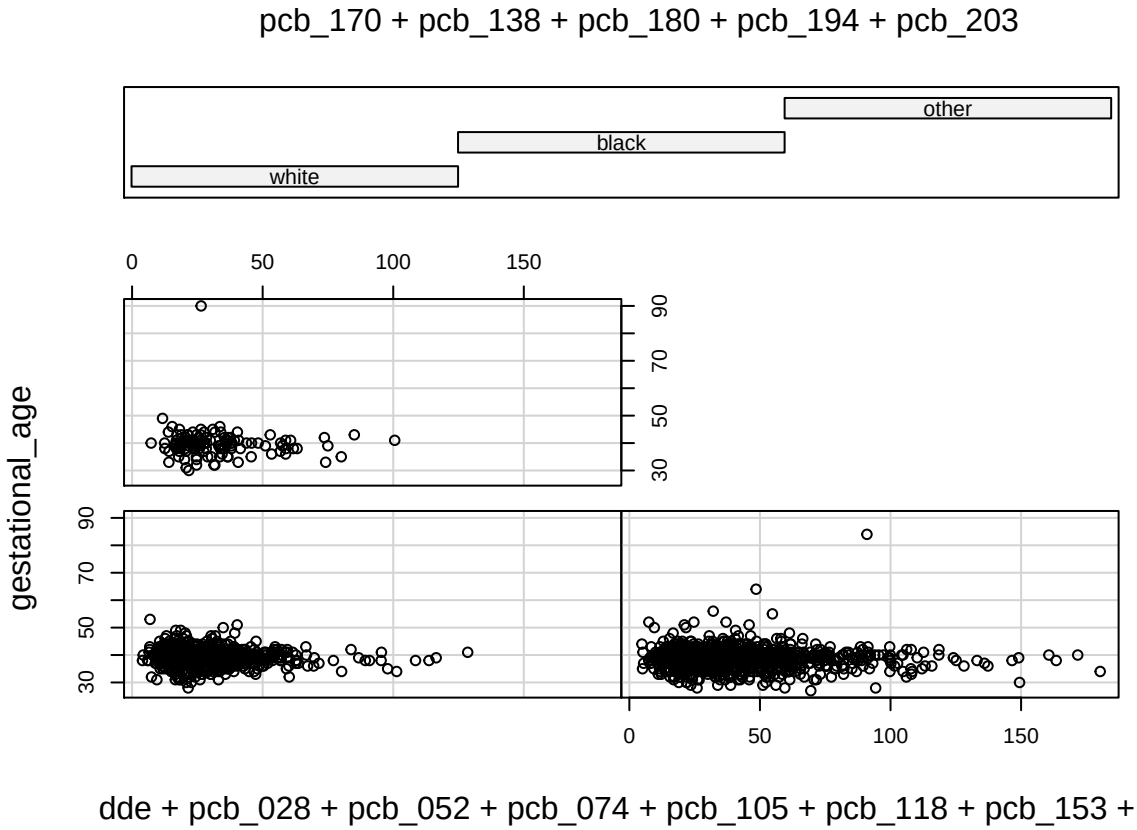

```
ggplot(dat, aes(group=smoking_status, x=smoking_status, y=gestational_age)) +
  geom_boxplot()
```



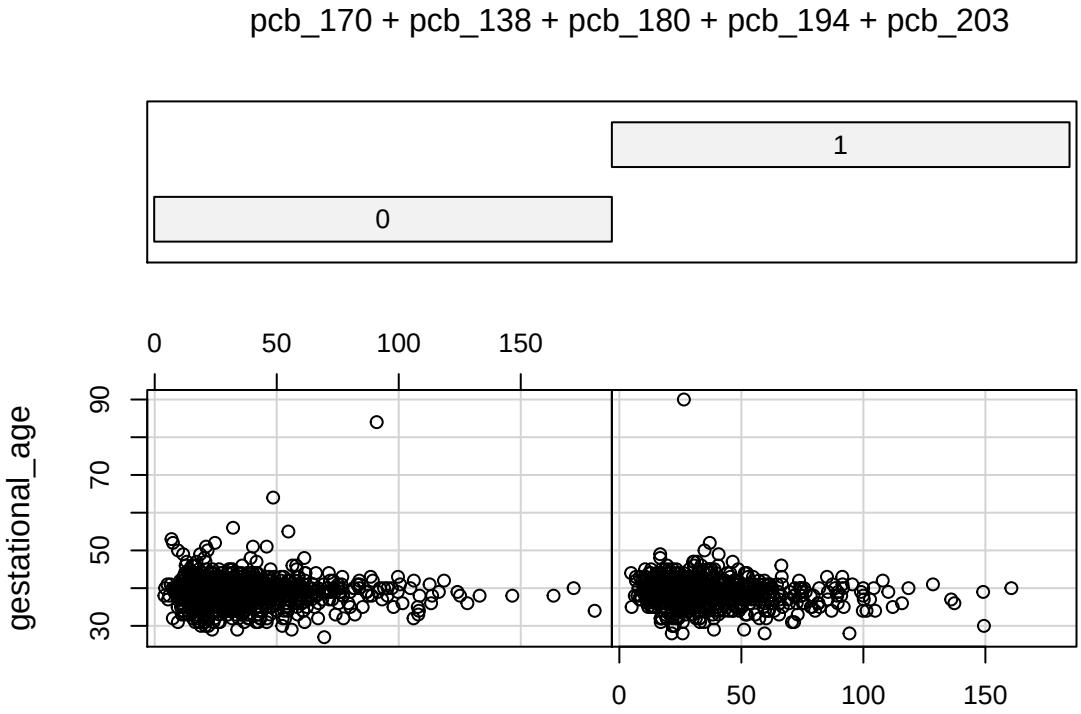

```
ggplot(dat, aes(group=center, x=center, y=gestational_age)) +
  geom_boxplot()
```



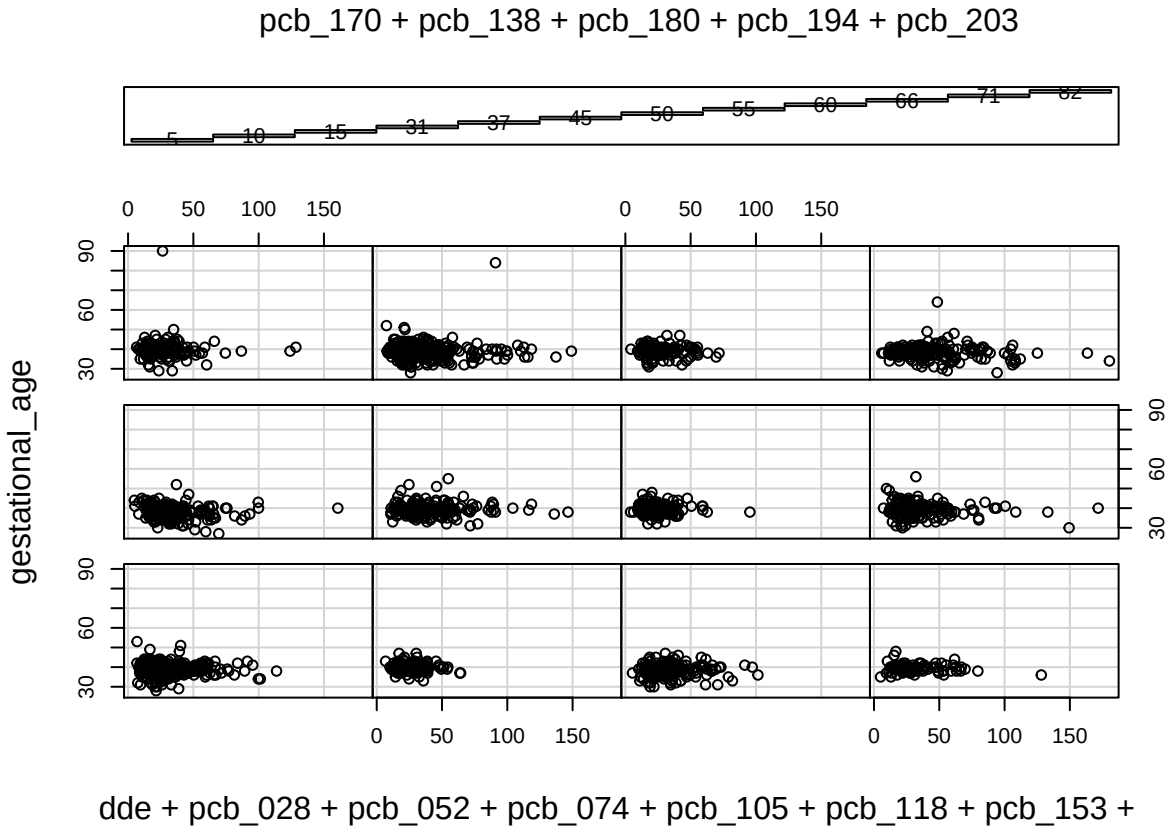

```
# Conditioning Plots
library(graphics)
coplot(gestational_age ~ dde + pcb_028 + pcb_052 + pcb_074 + pcb_105 + pcb_118 + pcb_153 + pcb_170 + pcb
```




```
coplot(gestational_age ~ dde + pcb_028 + pcb_052 + pcb_074 + pcb_105 + pcb_118 + pcb_153 + pcb_170 + pcb
```



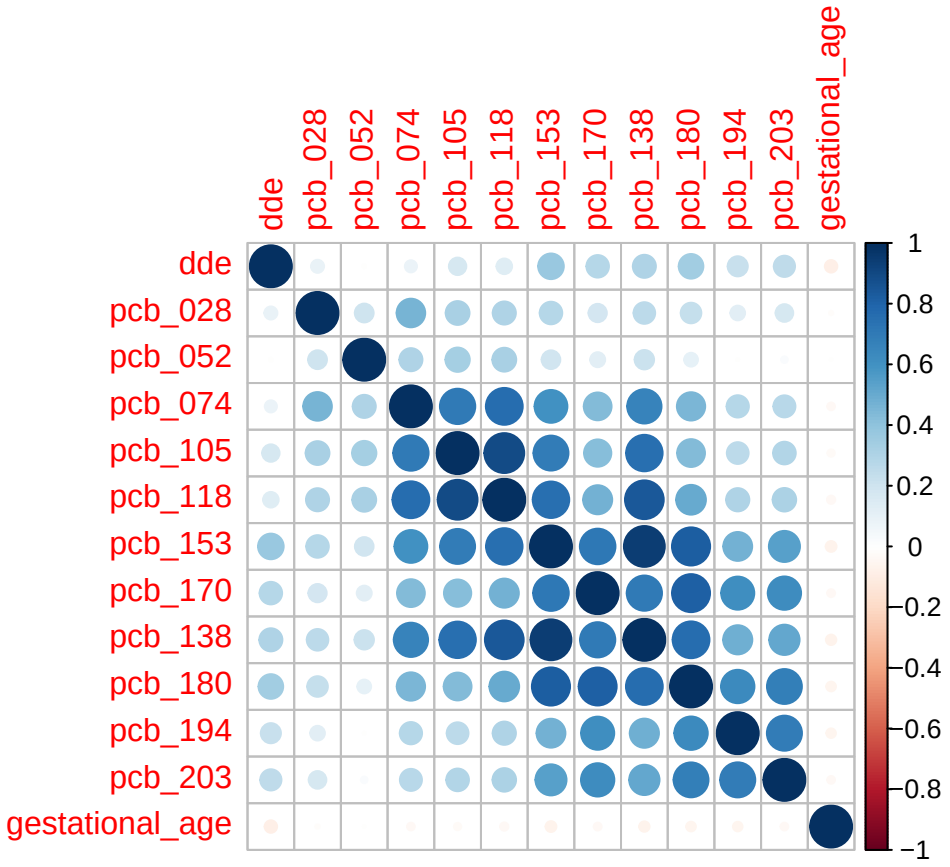

```
coplot(gestational_age ~ dde + pcb_028 + pcb_052 + pcb_074 + pcb_105 + pcb_118 + pcb_153 + pcb_170 + pcb
```




```
# Correlation Plot
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(dat[c(1:12,22)]))
```




```
# Summary Statistics
summary(dat)
```

```
##       dde              pcb_028          pcb_052           pcb_074
##  Min.   :  2.50   Min.   :0.000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.: 17.10   1st Qu.:0.110   1st Qu.:0.00000   1st Qu.:0.1500
##  Median : 24.70   Median :0.180   Median :0.00000   Median :0.2400
##  Mean   : 30.19   Mean   :0.195   Mean   :0.03053   Mean   :0.2692
##  3rd Qu.: 36.51   3rd Qu.:0.260   3rd Qu.:0.00000   3rd Qu.:0.3300
##  Max.   :178.06   Max.   :2.100   Max.   :1.80000   Max.   :2.4000
##
##     pcb_105          pcb_118          pcb_153          pcb_170
##  Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00   1st Qu.:0.3500   1st Qu.:0.4500   1st Qu.:0.00000
##  Median :0.11   Median :0.5400   Median :0.6300   Median :0.09000
##  Mean   :0.13   Mean   :0.6575   Mean   :0.7255   Mean   :0.09591
##  3rd Qu.:0.17   3rd Qu.:0.7900   3rd Qu.:0.8900   3rd Qu.:0.15000
##  Max.   :1.57   Max.   :5.6900   Max.   :3.5900   Max.   :0.59000
##
##     pcb_138          pcb_180          pcb_194          pcb_203
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.4100   1st Qu.:0.1400   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.5800   Median :0.2200   Median :0.00000   Median :0.00000
##  Mean   :0.6736   Mean   :0.2468   Mean   :0.04665   Mean   :0.04868
##  3rd Qu.:0.8300   3rd Qu.:0.3300   3rd Qu.:0.09000   3rd Qu.:0.10000
##  Max.   :3.8000   Max.   :1.2500   Max.   :0.40000   Max.   :0.59000
```

```
##
##     albumin      triglycerides     race      score_education  score_income
##  Min.   :2.600   Min.   :  51   white:1032   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:3.200   1st Qu.: 154   black:1223   1st Qu.:22.00   1st Qu.:22.00
##  Median :3.500   Median : 195   other: 124   Median :43.00   Median :52.00
##  Mean   :3.522   Mean   : 209                Mean   :48.96   Mean   :49.62
##  3rd Qu.:3.700   3rd Qu.: 247                3rd Qu.:64.00   3rd Qu.:75.00
##  Max.   :5.300   Max.   :1189                Max.   :97.00   Max.   :98.00
##  NA's   :2212                                NA's   :481     NA's   :515
##  score_occupation  maternal_age   smoking_status  cholesterol
##  Min.   : 0.00    Min.   :13.00   0:1327         Min.   : 55.0
##  1st Qu.:15.00    1st Qu.:20.00   1:1052         1st Qu.:195.0
##  Median :35.00    Median :23.00                  Median :232.0
##  Mean   :44.44    Mean   :24.24                  Mean   :237.5
##  3rd Qu.:73.00    3rd Qu.:28.00                  3rd Qu.:274.0
##  Max.   :94.00    Max.   :45.00                  Max.   :835.0
##  NA's   :479
##  gestational_age     center
##  Min.   :27.00   5      :485
##  1st Qu.:38.00   66     :395
##  Median :39.00   37     :207
##  Mean   :39.11   82     :192
##  3rd Qu.:41.00   45     :162
##  Max.   :90.00   15     :156
##                  (Other):782
```

Binary outcomes

```
dat0 <- dat
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
```

```
# Indicator for preterm
dat <- dat %>%
       mutate(ind_gest37 = if_else(gestational_age<37,1,0))

# Combine all pcb columns
pcb_col <- grep("pcb", names(dat))
dat <- dat %>%
  mutate(pcb_total = apply(dat[,pcb_col], 1, sum))

library(mice)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
apply(is.na(dat), 2, sum)
```

```
##              dde          pcb_028          pcb_052          pcb_074
##                0                0                0                0
##          pcb_105          pcb_118          pcb_153          pcb_170
##                0                0                0                0
##          pcb_138          pcb_180          pcb_194          pcb_203
##                0                0                0                0
##          albumin    triglycerides             race  score_education
##             2212                0                0              481
##     score_income score_occupation     maternal_age   smoking_status
##              515              479                0                0
##      cholesterol  gestational_age           center       ind_gest37
##                0                0                0                0
##        pcb_total
##                0
```

```
# remove albumin; impute score_education, score_income, score_occupation
dat_mice <- mice(dat[,-13], m=5, seed = 12345)
```

```
##
##  iter imp variable
##   1   1  score_education  score_income  score_occupation
##   1   2  score_education  score_income  score_occupation
##   1   3  score_education  score_income  score_occupation
##   1   4  score_education  score_income  score_occupation
##   1   5  score_education  score_income  score_occupation
##   2   1  score_education  score_income  score_occupation
##   2   2  score_education  score_income  score_occupation
##   2   3  score_education  score_income  score_occupation
##   2   4  score_education  score_income  score_occupation
##   2   5  score_education  score_income  score_occupation
##   3   1  score_education  score_income  score_occupation
##   3   2  score_education  score_income  score_occupation
##   3   3  score_education  score_income  score_occupation
##   3   4  score_education  score_income  score_occupation
##   3   5  score_education  score_income  score_occupation
##   4   1  score_education  score_income  score_occupation
##   4   2  score_education  score_income  score_occupation
##   4   3  score_education  score_income  score_occupation
##   4   4  score_education  score_income  score_occupation
##   4   5  score_education  score_income  score_occupation
##   5   1  score_education  score_income  score_occupation
```

```
##   5   2  score_education  score_income  score_occupation
##   5   3  score_education  score_income  score_occupation
##   5   4  score_education  score_income  score_occupation
##   5   5  score_education  score_income  score_occupation
```

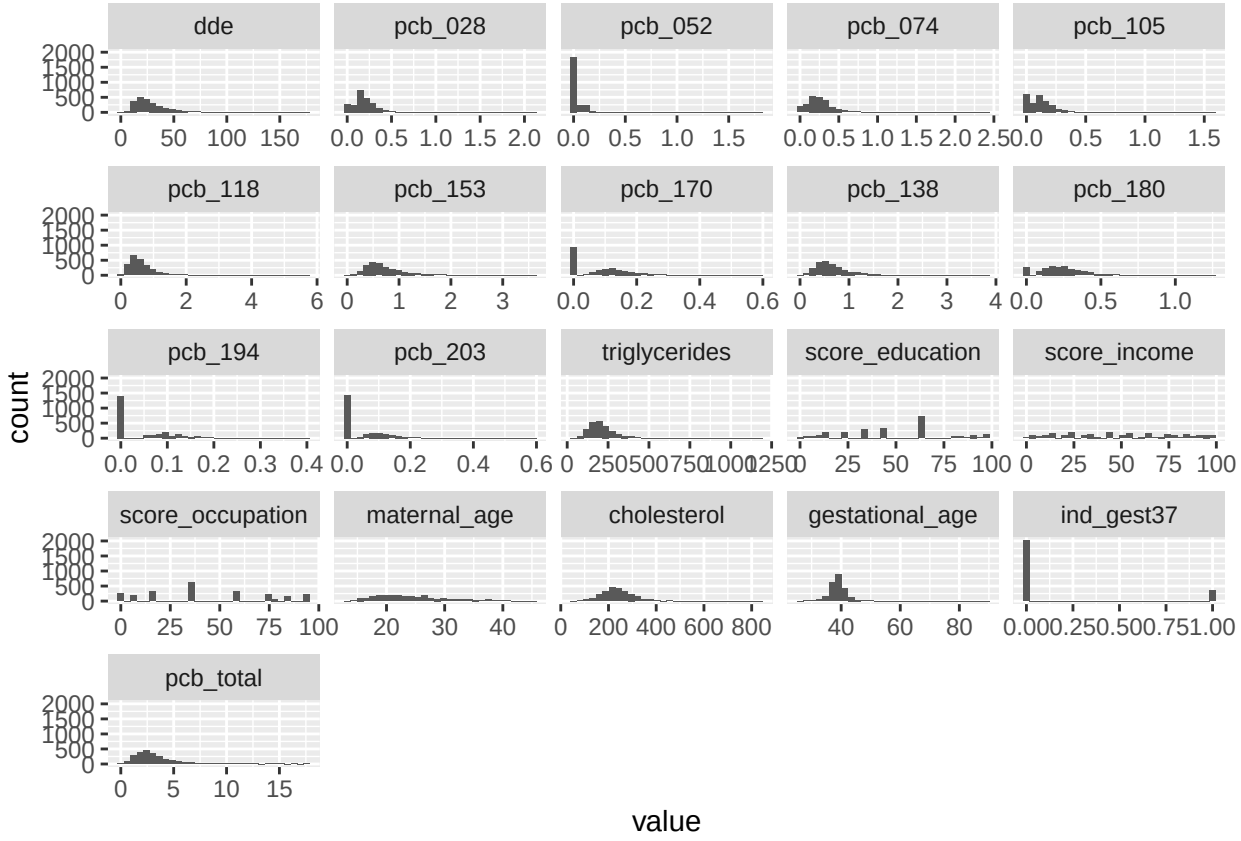
```
## Warning: Number of logged events: 75
```

```
# complete(dat_mice)

# Further EDA
d <- melt(complete(dat_mice))
```

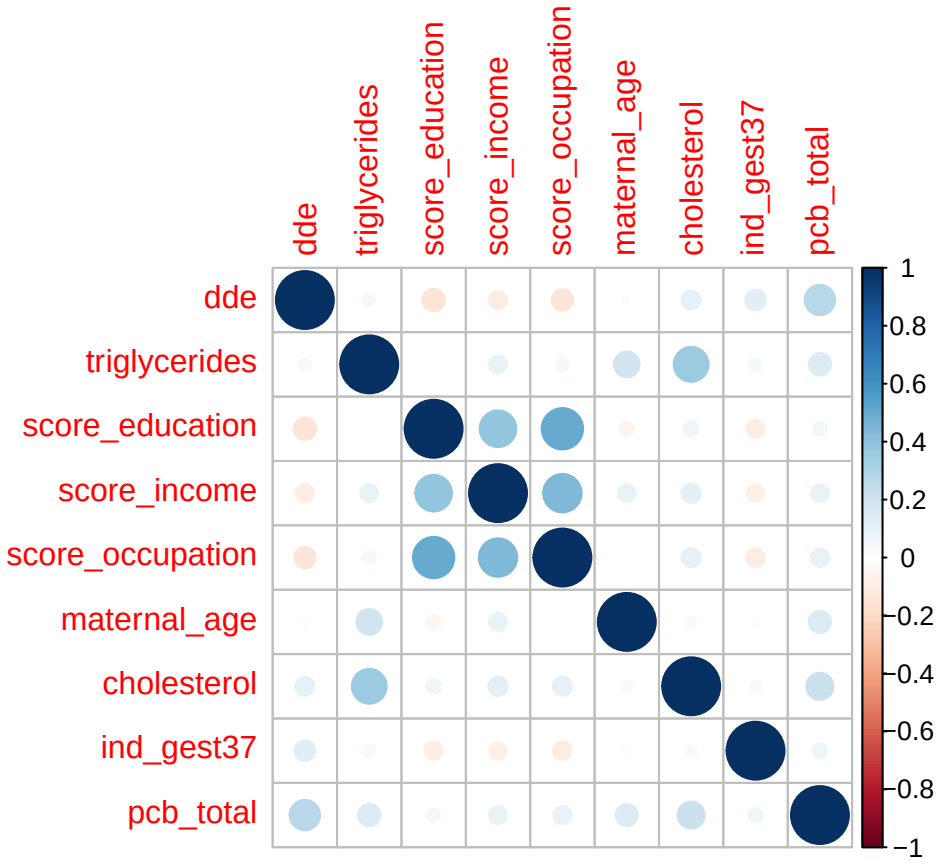
```
## Using race, smoking_status, center as id variables
```

```
ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
corrplot(cor(complete(dat_mice)[,c(1,13,15:18,20,23:24)]))
```




```
pcb_col <- grep("pcb", names(dat0))
pcb_colnames <- paste(colnames(dat0)[pcb_col], collapse = "+",sep = "")
confound_colnames <- paste(colnames(dat0[c(14:21,23)]),collapse = "+", sep = "")
full_formula_ind <- as.formula(paste("ind_gest37~dde+",pcb_colnames, "+", confound_colnames, seq = ""))
# fit2 <- glm(full_formula_ind, data = dat, family = "binomial"(link="logit"))
fit1 <- with(data = dat_mice, exp = glm(ind_gest37 ~ dde + pcb_028 + pcb_052 + pcb_074 + pcb_105 + pcb_
    triglycerides + race + score_education + score_income + score_occupation +
    maternal_age + smoking_status + cholesterol + center, family  = "binomial"(link="logit")))
fit1_pool <- pool(fit1)
summary(fit1_pool)
```

```
##                    estimate    std.error   statistic         df      p.value
## (Intercept)      -2.239755968 0.4719387884 -4.74586116 2041.99229 2.220193e-06
## dde               0.007490606 0.0031118572  2.40711755 2336.58369 1.615610e-02
## pcb_028          -0.126581351 0.4815729206 -0.26284981 2340.35005 7.926895e-01
## pcb_052           0.509018177 0.7168979168  0.71002881 2340.34526 4.777570e-01
## pcb_074           0.734812378 0.4517821521  1.62647501 2343.62843 1.039831e-01
## pcb_105           0.021300182 1.0334150247  0.02061145 2342.86078 9.835574e-01
## pcb_118          -0.343204070 0.4026407153 -0.85238292 2341.97938 3.940888e-01
## pcb_153           0.465705447 0.5357793240  0.86921131 2324.34315 3.848213e-01
## pcb_170          -1.822419279 1.1089067905 -1.64343775 2331.46417 1.004272e-01
## pcb_138           0.305790751 0.6642411201  0.46036107 2339.87581 6.452999e-01
## pcb_180           0.237309610 0.8250023824  0.28764718 2334.22582 7.736424e-01
## pcb_194           0.066119962 1.4551041269  0.04544002 2342.67402 9.637605e-01
## pcb_203           0.789333584 1.3405109137  0.58883040 2335.93487 5.560320e-01
```

```
## triglycerides     0.003129335 0.0007868986  3.97679636 2343.70817 7.196654e-05
## raceblack         0.188839631 0.2167796416  0.87111331 2312.28039 3.837827e-01
## raceother         0.435603732 0.3544212907  1.22905633 2333.34254 2.191746e-01
## score_education  -0.003027195 0.0028224103 -1.07255682  686.84343 2.838467e-01
## score_income     -0.002161175 0.0029252033 -0.73881189   42.67258 4.640657e-01
## score_occupation -0.002507316 0.0025958860 -0.96588069  351.43334 3.347681e-01
## maternal_age     -0.013940566 0.0104642108 -1.33221379 2262.37222 1.829241e-01
## smoking_status1   0.119949638 0.1253733922  0.95673919 2334.18475 3.387980e-01
## cholesterol      -0.002486336 0.0010208094 -2.43565141 2335.71180 1.493933e-02
## center10         -0.977967583 0.4938667403 -1.98022564 2343.48096 4.779497e-02
## center15          0.759236804 0.3370392947  2.25266554 1575.14139 2.441740e-02
## center31         -0.592244818 0.4912285156 -1.20564014 2340.46336 2.280781e-01
## center37          0.766897505 0.2828718986  2.71111238 2306.87144 6.755490e-03
## center45          0.085130009 0.3313721578  0.25690151 1978.93291 7.972815e-01
## center50         -0.041690474 0.3515749119 -0.11858205 2338.40269 9.056167e-01
## center55          0.313319819 0.3631964624  0.86267310 2229.59634 3.884100e-01
## center60          0.321783030 0.3232783089  0.99537464 2277.37226 3.196598e-01
## center66          0.202801829 0.2785524753  0.72805610 2248.59049 4.666551e-01
## center71         -0.085236955 0.3198593331 -0.26648262 2278.40847 7.898917e-01
## center82          0.549194229 0.3278861155  1.67495421 2101.88244 9.409197e-02
```

```
fit2 <- glm(full_formula_ind, data = complete(dat_mice),
            family = "binomial"(link="logit"))
dat2 <- complete(dat_mice) %>%
        mutate(Residuals = residuals.glm(fit2,type="response"),
         Predicted = predict.glm(fit2,type="response"))
library(arm)
```

```
## Loading required package: MASS
```

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

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

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

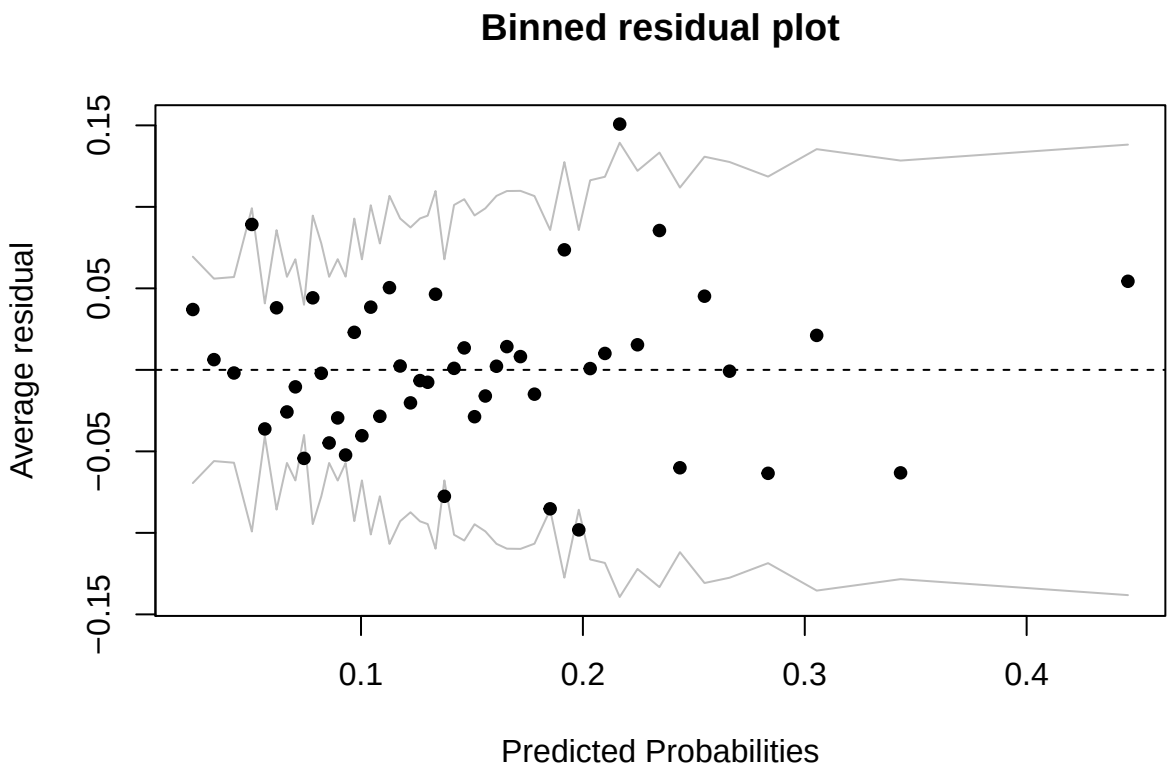
```
##
## arm (Version 1.10-1, built: 2018-4-12)
```

```
## Working directory is /Users/yiji/Desktop/Duke Statistics PhD/2020 Spring Courses/STA 723 Case Studies
```

```
##
## Attaching package: 'arm'
```

```
## The following object is masked from 'package:corrplot':
##
##     corrplot
```

```
binnedplot(x=dat2$Predicted,y=dat2$Residuals,
           xlab="Predicted Probabilities")
```

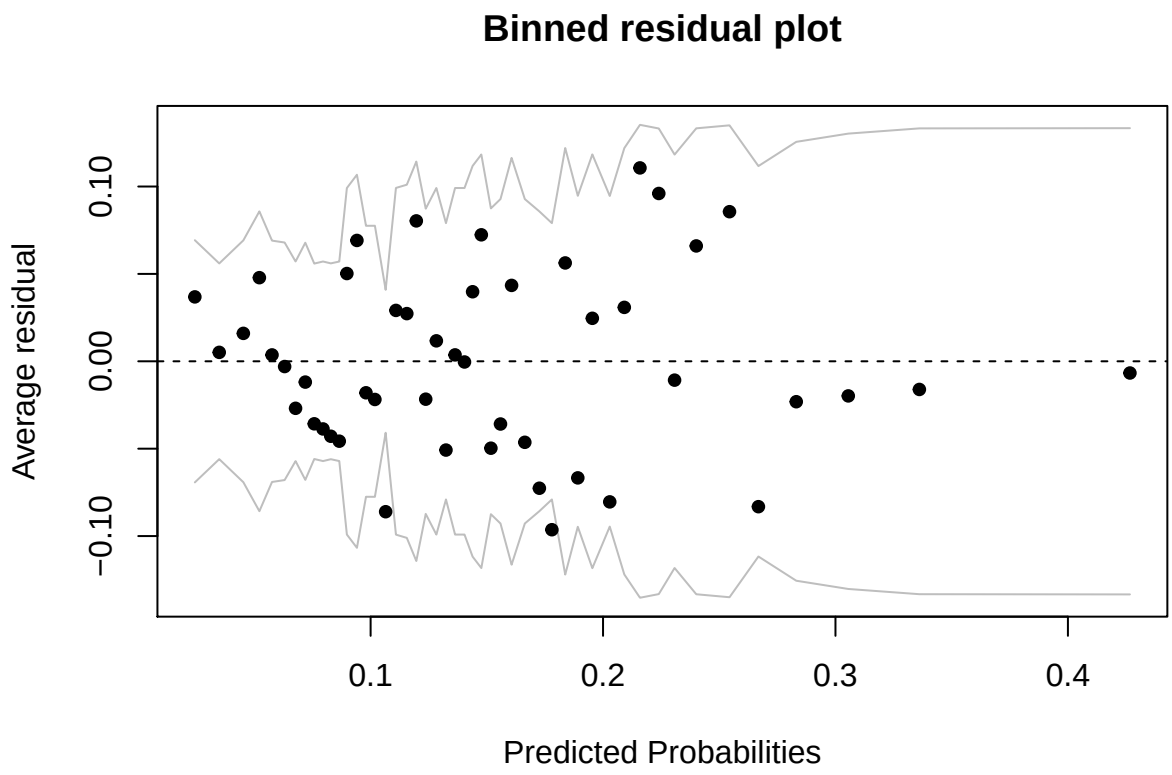



```
fit1.1 <- with(data = dat_mice, exp = glm(ind_gest37 ~ dde + pcb_total +
    triglycerides + race + score_education + score_income + score_occupation +
    maternal_age + smoking_status + cholesterol + center,
    family = "binomial"(link="logit")))
fit1.1_pool <- pool(fit1.1)
summary(fit1.1_pool)
```

```
##                     estimate    std.error  statistic         df      p.value
## (Intercept)      -2.233661833 0.4527310518 -4.9337500 2133.61921 8.690376e-07
## dde               0.008354011 0.0029340745  2.8472390 2341.32573 4.448463e-03
## pcb_total         0.107182020 0.0355175685  3.0177184 2348.15276 2.574220e-03
## triglycerides     0.003025333 0.0007760362  3.8984432 2353.75329 9.952008e-05
## raceblack         0.217699769 0.2117999856  1.0278554 2313.93649 3.041253e-01
## raceother         0.426033532 0.3525113865  1.2085667 2337.34617 2.269516e-01
## score_education  -0.003149136 0.0027970767 -1.1258669  764.76955 2.605750e-01
## score_income     -0.001964270 0.0028766801 -0.6828252   47.04292 4.980663e-01
## score_occupation -0.002566643 0.0025893157 -0.9912439  340.48404 3.222704e-01
## maternal_age     -0.012675984 0.0100795515 -1.2575941 2241.50900 2.086696e-01
## smoking_status1   0.156446431 0.1221493965  1.2807794 2341.29517 2.003980e-01
## cholesterol      -0.002449080 0.0010053749 -2.4359865 2343.96670 1.492527e-02
## center10         -0.942703324 0.4884588607 -1.9299544 2353.54627 5.373253e-02
## center15          0.812261072 0.3293132278  2.4665304 1742.92197 1.373851e-02
```

```
## center31         -0.574675217 0.4850223533 -1.1848427 2350.35405 2.361994e-01
## center37          0.772226679 0.2658955417  2.9042483 2317.35271 3.716204e-03
## center45          0.113280953 0.3175432039  0.3567419 2054.29550 7.213217e-01
## center50          0.058826292 0.3387395831  0.1736623 2352.37884 8.621458e-01
## center55          0.365345112 0.3554367403  1.0278766 2246.13615 3.041185e-01
## center60          0.384293296 0.3160174231  1.2160510 2305.48688 2.240900e-01
## center66          0.229983177 0.2694263627  0.8536031 2271.21659 3.934150e-01
## center71         -0.045225559 0.3103564697 -0.1457213 2297.77329 8.841542e-01
## center82          0.518379911 0.3212681282  1.6135429 2119.22052 1.067755e-01
```

```
fit2.1 <- glm(ind_gest37 ~ dde + pcb_total +
    triglycerides + race + score_education + score_income + score_occupation +
    maternal_age + smoking_status + cholesterol + center, data = complete(dat_mice),
            family = "binomial"(link="logit"))
dat2.1 <- complete(dat_mice) %>%
        mutate(Residuals = residuals.glm(fit2.1,type="response"),
         Predicted = predict.glm(fit2.1,type="response"))
library(arm)
binnedplot(x=dat2.1$Predicted,y=dat2.1$Residuals,
           xlab="Predicted Probabilities")
```



Multi-level outcomes

```
dat0 <- dat
library(dplyr)
# Indicator for preterm
dat <- dat %>%
       mutate(preterm_ind = if_else(gestational_age<33,2,
                                    if_else(gestational_age<37 &
                                              gestational_age>32,1,0)))

# Combine all pcb columns
pcb_col <- grep("pcb", names(dat))
dat <- dat %>%
  mutate(pcb_total = apply(dat[,pcb_col], 1, sum))

library(mice)
apply(is.na(dat), 2, sum)
```

```
##              dde          pcb_028          pcb_052          pcb_074
##                0                0                0                0
##          pcb_105          pcb_118          pcb_153          pcb_170
##                0                0                0                0
##          pcb_138          pcb_180          pcb_194          pcb_203
##                0                0                0                0
##          albumin    triglycerides             race  score_education
##             2212                0                0              481
##     score_income score_occupation     maternal_age   smoking_status
##              515              479                0                0
##      cholesterol  gestational_age           center       ind_gest37
##                0                0                0                0
##        pcb_total      preterm_ind
##                0                0
```

```
# remove albumin; impute score_education, score_income, score_occupation
dat_mice <- mice(dat[,-13], m=5, seed = 12345)
```

```
##
##  iter imp variable
##   1   1  score_education  score_income  score_occupation
##   1   2  score_education  score_income  score_occupation
##   1   3  score_education  score_income  score_occupation
##   1   4  score_education  score_income  score_occupation
##   1   5  score_education  score_income  score_occupation
##   2   1  score_education  score_income  score_occupation
##   2   2  score_education  score_income  score_occupation
##   2   3  score_education  score_income  score_occupation
##   2   4  score_education  score_income  score_occupation
##   2   5  score_education  score_income  score_occupation
##   3   1  score_education  score_income  score_occupation
##   3   2  score_education  score_income  score_occupation
##   3   3  score_education  score_income  score_occupation
##   3   4  score_education  score_income  score_occupation
##   3   5  score_education  score_income  score_occupation
##   4   1  score_education  score_income  score_occupation
```

```
##   4   2  score_education  score_income  score_occupation
##   4   3  score_education  score_income  score_occupation
##   4   4  score_education  score_income  score_occupation
##   4   5  score_education  score_income  score_occupation
##   5   1  score_education  score_income  score_occupation
##   5   2  score_education  score_income  score_occupation
##   5   3  score_education  score_income  score_occupation
##   5   4  score_education  score_income  score_occupation
##   5   5  score_education  score_income  score_occupation
```

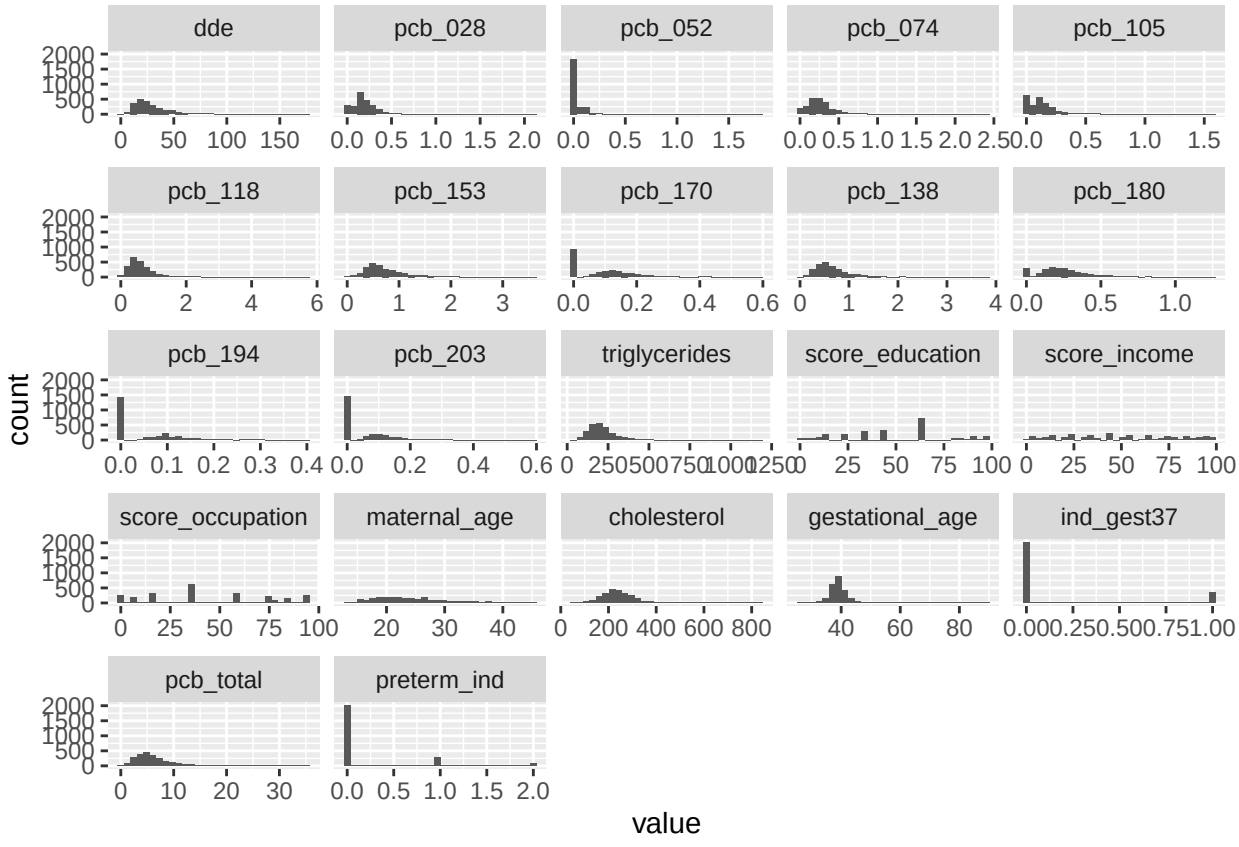
```
## Warning: Number of logged events: 75
```

```
imp_dat <- complete(dat_mice)

# Further EDA
d <- melt(imp_dat)
```

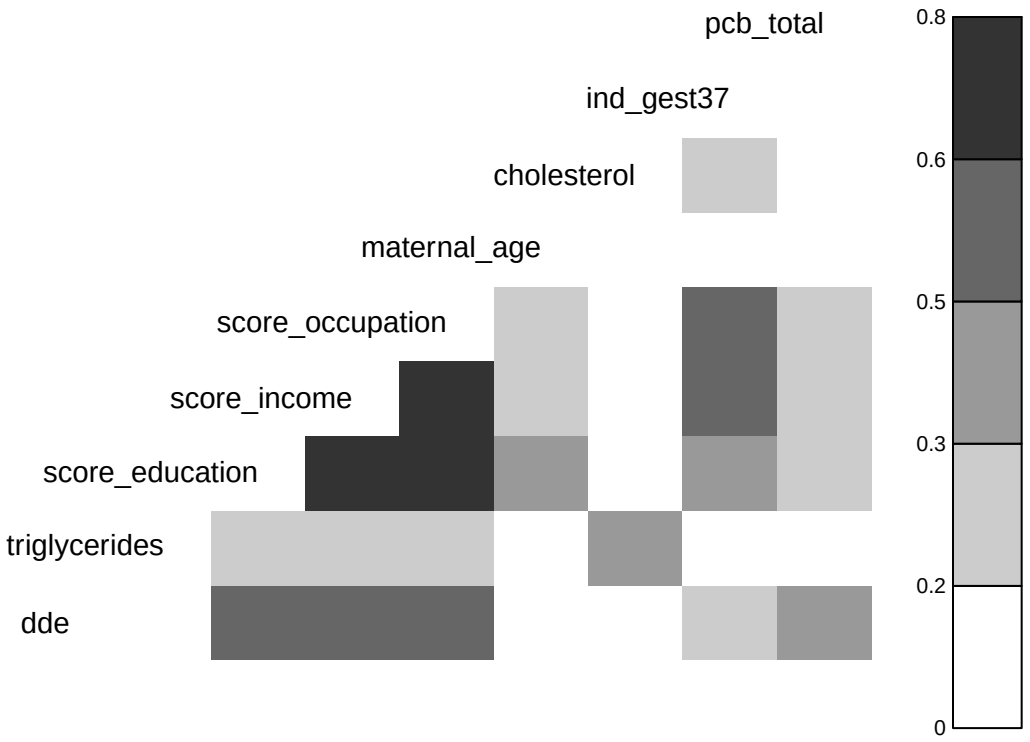
```
## Using race, smoking_status, center as id variables
```

```
ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
corrplot(cor(imp_dat[,c(1,13,15:18,20,23:24)]))
```



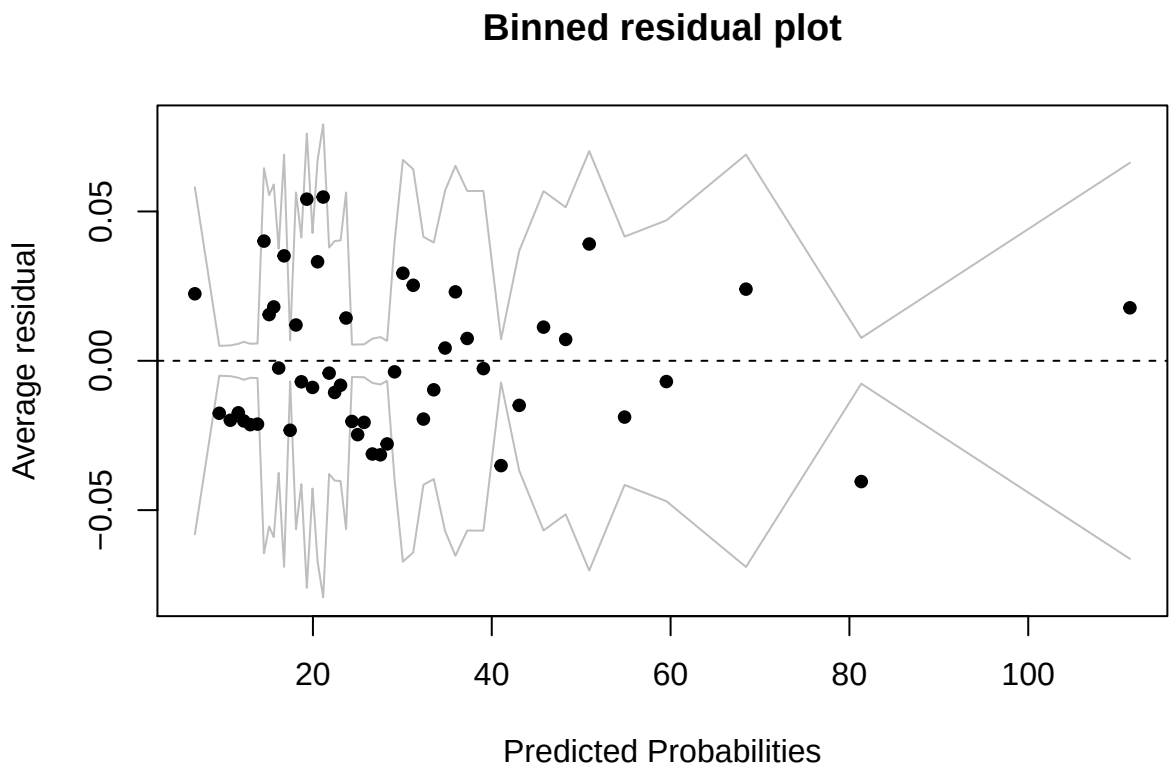

```
pcb_col <- grep("pcb", names(dat0))
pcb_colnames <- paste(colnames(dat0)[pcb_col], collapse = "+",sep = "")
confound_colnames <- paste(colnames(dat0[c(14:21,23)]),collapse = "+", sep = "")
full_formula_ind <- as.formula(paste("preterm_ind~dde+",pcb_colnames, "+", confound_colnames, seq = ""))

library(nnet)
library(broom)
fit1 <- multinom(full_formula_ind, data = imp_dat)
```

```
## # weights:  105 (68 variable)
## initial  value 2613.598635
## iter  10 value 1646.456174
## iter  20 value 1440.142645
## iter  30 value 1196.975770
## iter  40 value 1107.114988
## iter  50 value 1105.848056
## iter  60 value 1105.834184
## final  value 1105.833634
## converged
```

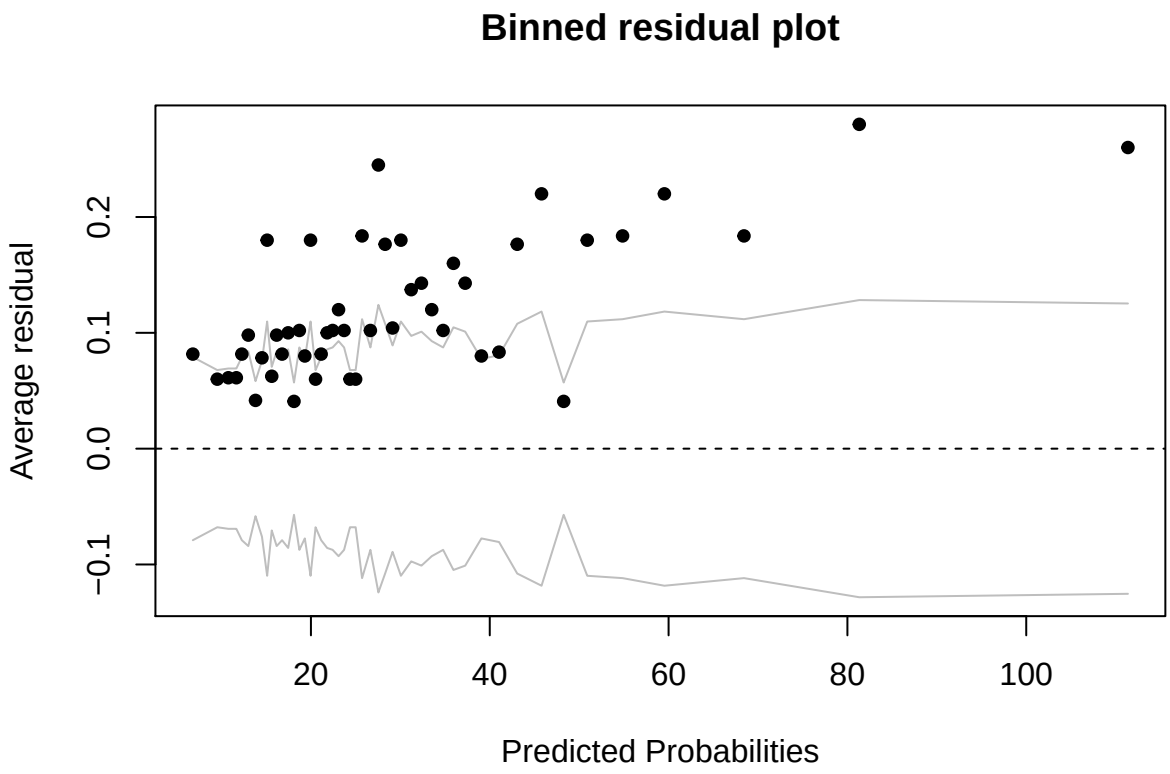
```
tidy(fit1,exponentiate=FALSE) #display log-odds model
```

```
## # A tibble: 68 x 6
##    y.level term        estimate std.error statistic     p.value
##    <chr>   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
##  1 1       (Intercept) -2.69      0.508     -5.29   0.000000121
##  2 1       dde          0.00832   0.00329    2.53   0.0115
##  3 1       pcb_028     -0.354     0.538     -0.658  0.510
##  4 1       pcb_052      0.428     0.734      0.583  0.560
##  5 1       pcb_074      0.596     0.513      1.16   0.246
##  6 1       pcb_105      0.683     1.06       0.645  0.519
##  7 1       pcb_118     -0.645     0.514     -1.25   0.210
##  8 1       pcb_153      0.539     0.603      0.893  0.372
##  9 1       pcb_170     -1.66      1.16      -1.43   0.154
## 10 1       pcb_138      0.0238    0.734      0.0324 0.974
## # ... with 58 more rows
```

```
# calculate predicted probabilities
pred.probs <- predict(fit1,type="probs")

# calculate residuals for category j
very_preterm <- if_else(imp_dat$preterm_ind==2,1,0)
residual_very_preterm <- very_preterm - pred.probs[,3]

preterm <- if_else(imp_dat$preterm_ind==1,1,0)
residual_preterm <- preterm - pred.probs[,2]

library(arm)
binnedplot(x=imp_dat$dde,y=residual_very_preterm,
           xlab="Predicted Probabilities")
```




```
binnedplot(x=imp_dat$dde,y=preterm,
           xlab="Predicted Probabilities")
```




```
fit2 <- multinom(preterm_ind ~ dde + pcb_total + triglycerides + race + score_education + score_income +
center, data = imp_dat)
```

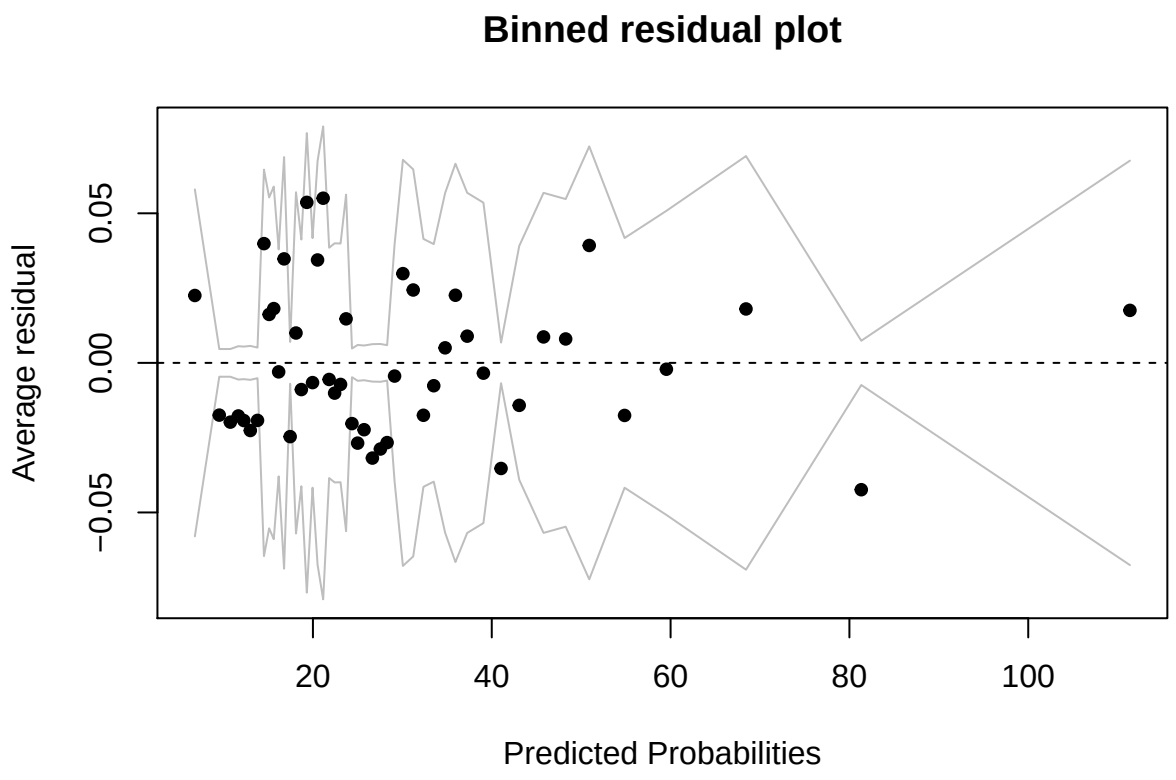
```
## # weights:  72 (46 variable)
## initial  value 2613.598635
## iter  10 value 1646.474774
## iter  20 value 1425.439049
## iter  30 value 1184.777094
## iter  40 value 1112.982399
## iter  50 value 1112.343109
## final  value 1112.342927
## converged
```

```
tidy(fit2,exponentiate=FALSE) #display log-odds model
```

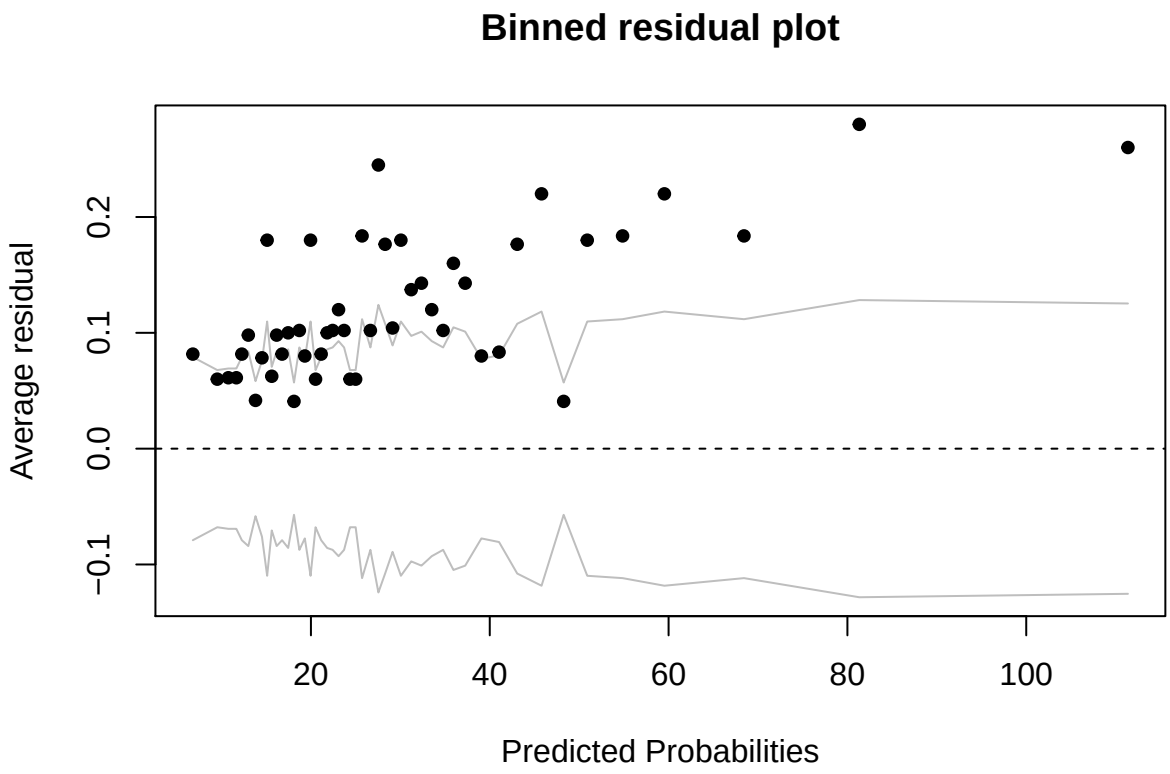
```
## # A tibble: 46 x 6
##    y.level term             estimate std.error statistic      p.value
##    <chr>   <chr>               <dbl>     <dbl>     <dbl>        <dbl>
##  1 1       (Intercept)      -2.71     0.492       -5.52  0.0000000343
##  2 1       dde               0.00955  0.00312      3.07  0.00217
##  3 1       pcb_total         0.0454   0.0195       2.32  0.0201
##  4 1       triglycerides     0.00311  0.000827     3.76  0.000170
##  5 1       raceblack         0.0522   0.229        0.228 0.819
##  6 1       raceother         0.144    0.386        0.374 0.708
```

```
##  7 1       score_education  -0.00379  0.00295     -1.28  0.199
##  8 1       score_income      0.00237  0.00261      0.909 0.363
##  9 1       score_occupation -0.00290  0.00268     -1.08  0.280
## 10 1       maternal_age     -0.0140   0.0110      -1.27  0.203
## # ... with 36 more rows
```

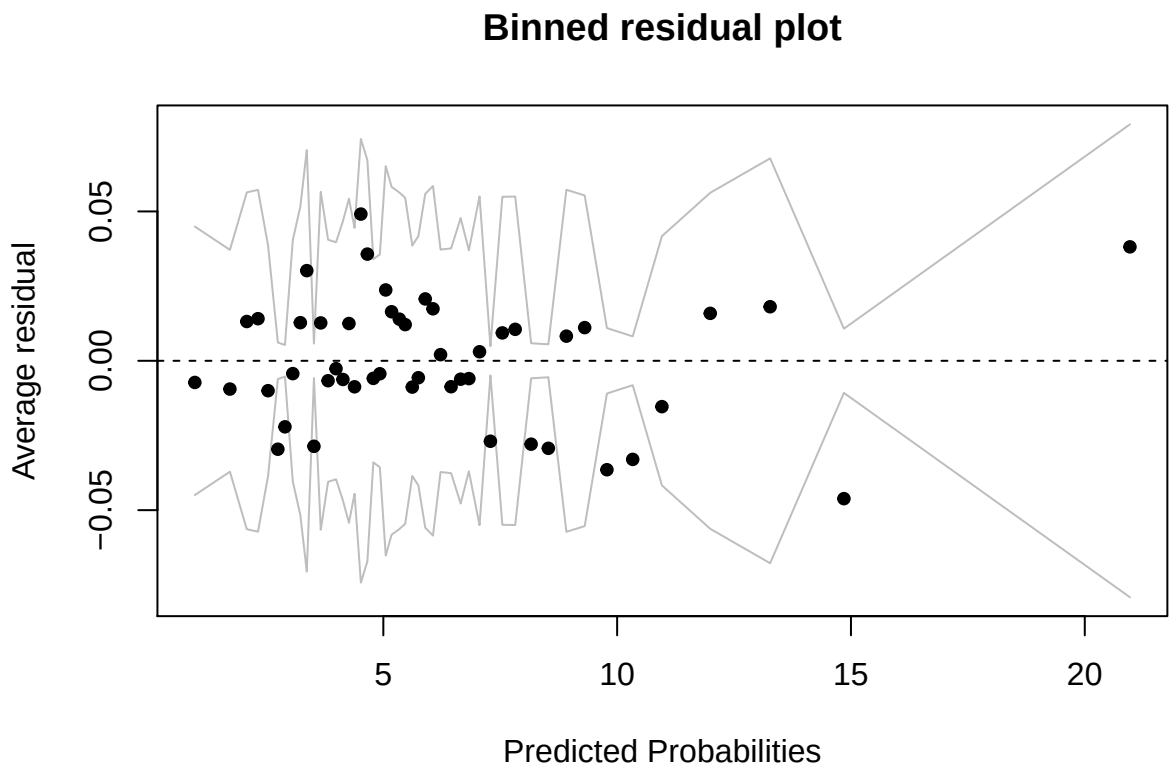
```
# calculate predicted probabilities
pred.probs <- predict(fit2,type="probs")

# calculate residuals for category j
very_preterm <- if_else(imp_dat$preterm_ind==2,1,0)
residual_very_preterm <- very_preterm - pred.probs[,3]

preterm <- if_else(imp_dat$preterm_ind==1,1,0)
residual_preterm <- preterm - pred.probs[,2]

library(arm)
binnedplot(x=imp_dat$dde,y=residual_very_preterm,
           xlab="Predicted Probabilities")
```



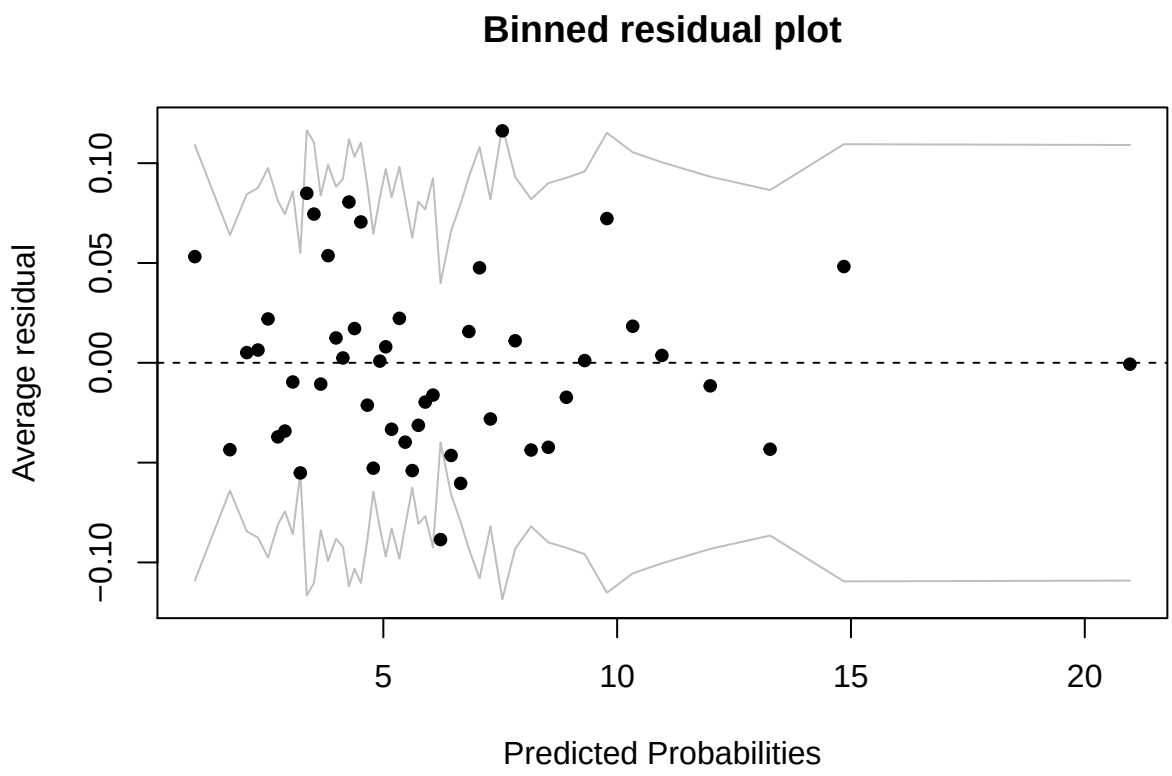

```
binnedplot(x=imp_dat$dde,y=preterm,
           xlab="Predicted Probabilities") # residual plot!!
```




```
binnedplot(x=imp_dat$pcb_total,y=residual_very_preterm,
           xlab="Predicted Probabilities")
```




```
binnedplot(x=imp_dat$pcb_total,y=residual_preterm,
           xlab="Predicted Probabilities")
```




```
anova(fit1, fit2, test="Chisq")
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: preterm_ind
##
## 1
## 2 dde + pcb_028 + pcb_052 + pcb_074 + pcb_105 + pcb_118 + pcb_153 + pcb_170 + pcb_138 + pcb_180 + pcb
##   Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1      4712   2224.686
## 2      4692   2211.667 1 vs 2    20 13.01859 0.8765852
```

try ordinal logistic regression? (polr)

```
# # Extract PCB columns & Create full formula
# pcb_col <- grep("pcb", names(dat))
# pcb_colnames <- paste(colnames(dat)[pcb_col], collapse = "+",sep = "")
# confound_colnames <- paste(colnames(dat[c(14:21,23)]),collapse = "+", sep = "")
# full_formula <- as.formula(paste("gestational_age~dde+",pcb_colnames, "+", confound_colnames, seq = "

# # Remove pcb138 (collinearity)
# pcb_colnames_no138 <- gsub("\\+pcb_138", "", pcb_colnames)
# full_formula_no138 <- as.formula(paste("gestational_age~dde+",pcb_colnames_no138, "+", confound_colna
```