Case Study #1

Shrey Gupta, Frances Hung, Ezinne Nwankwo

0: Abstract

1: Introduction

We study how DDE (Dichlorodiphenyldichloroethylene) and PCBs (Polychlorinated Biphenyls) relate to risk of premature delivery, which is associated with higher risk of morbidity and mortality for the newborn. We use a sample of 2,380 women and children from Longnecker, et al. (2001) initially provided by the National Collaborative Perinatal Project. DDE and PCBs have been used to treat crops in order to limit their predation, and, as a result of their non-biodegradability, remain present in environments where humans can be exposed to them. These chemicals build up in fat in human tissues, and can have an impact on human health, including risk of premature delivery.

The data include various demographic variables (race, age, and socioeconomic index), smoking status, concentration doses of DDE and PCBs due to exposure, and cholesterol and triglycerides levels. We define pre-term pregnancy with a cut-off of 36 weeks or fewer, which tends to be the region around which there begins higher risk of morbidity and mortality for the child.

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.

Since a substantial portion of observations didn't have an albumin measurement, we disregarded it in our analysis. Due to the lack of interpretability and missingness of the scores for education, occupation, and income, we also didn't consider using these variables in our model. Getting rid of these covariates in our analysis removed most of the missingness from our dataset.

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 logistic model with a random-effect intercept:

We evaluate model fit using BIC and AIC.

3: Results

3.1: Exploratory Data Analysis

3.2: Main Results

3.3: Sensitivity Analysis

4: Discussion

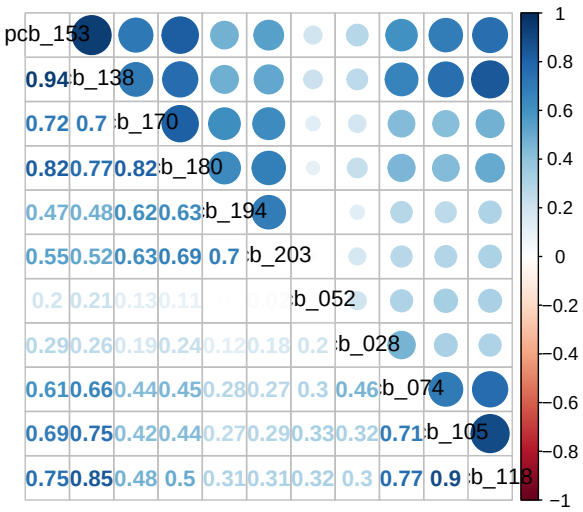
Our results find that the effect for average PCB is not significant, meaning there is no evidence of an association between PCB exposure and risk of pre-term delivery. However, higher DDE exposure is associated with higher risk of pre-term delivery. A one unit increase in DDE is associated with decreasing 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.

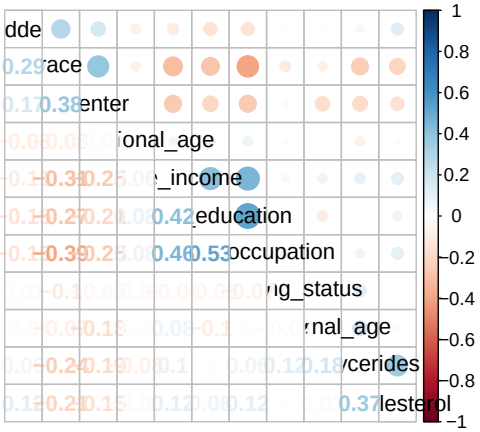
There are various advantages and disadvantages of the approach we took. On one hand, regression is highly interpretable, and interpretability is important for disciplines like the health and sciences. (We also used the min, max, and average PCB exposures instead of doing PCA since the former is more 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

```
par(mfrow=c(1,2))
corrplot.mixed(cor(data %>% select(starts_with("pcb_")) %>% drop_na() %>% sapply(., as.numeric)),
               order="hclust", tl.col="black")
corrplot.mixed(cor(data %>% select(-starts_with("pcb_"), -albumin) %>% drop_na() %>%
                     sapply(., as.numeric)), order="hclust", tl.col="black")
```

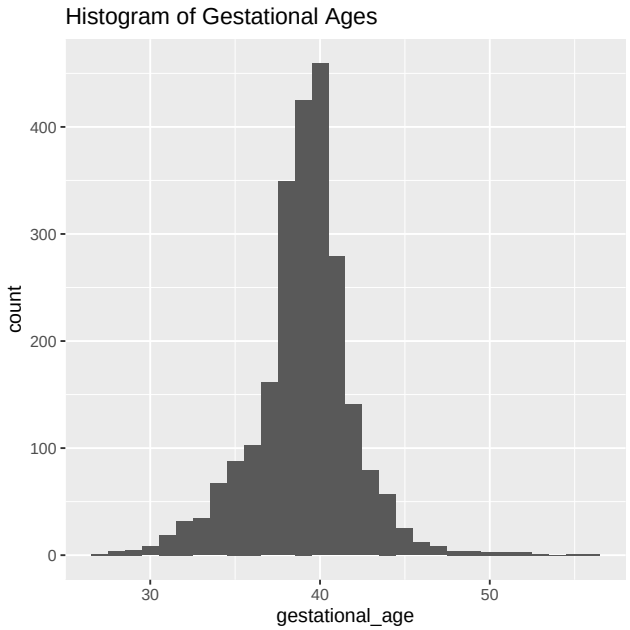


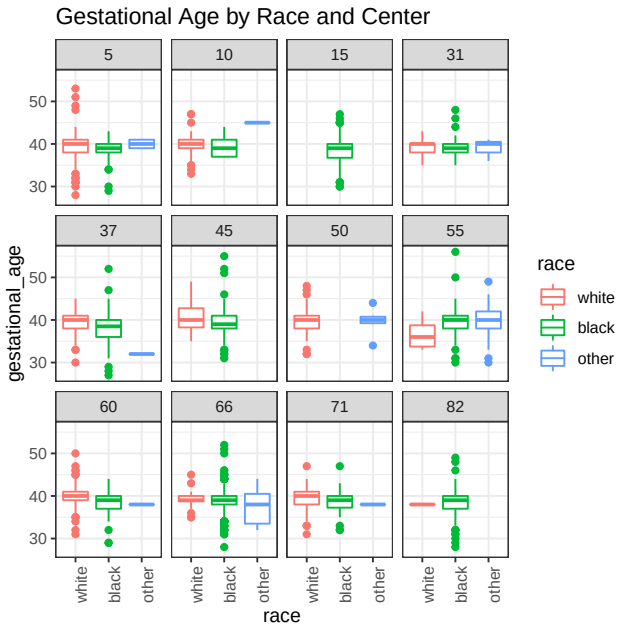


We observe that PCB variations are positively correlated with one another, and that certain groups of variables are also correlated (education, occupation, and income; triglycerides and cholesterol; race and center; race and DDE; maternal age and triglycerides, etc.).

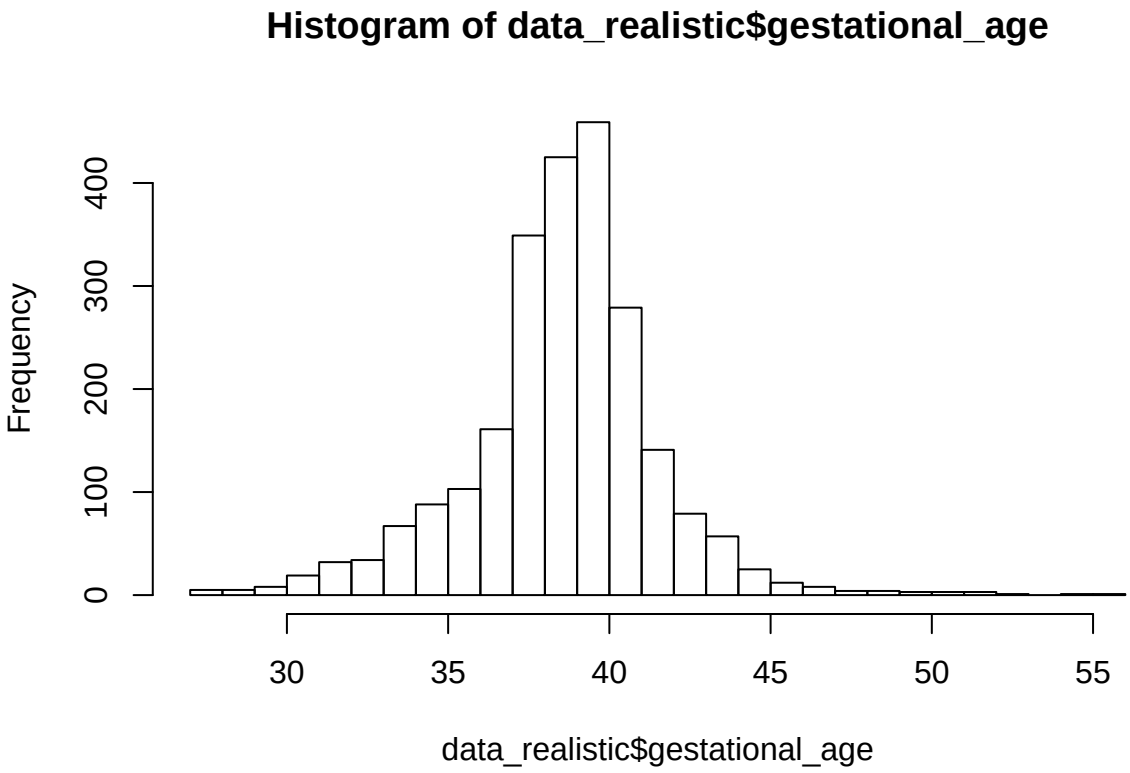
```
## # A tibble: 31 x 3
## # Groups:   race [3]
##    race  center n_race
##    <fct> <fct>   <int>
##  1 white 5         431
##  2 white 10        122
##  3 white 31         21
##  4 white 37         46
##  5 white 45         30
##  6 white 50        141
##  7 white 55          8
##  8 white 60         86
##  9 white 66         28
## 10 white 71        118
## # ... with 21 more rows
```

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



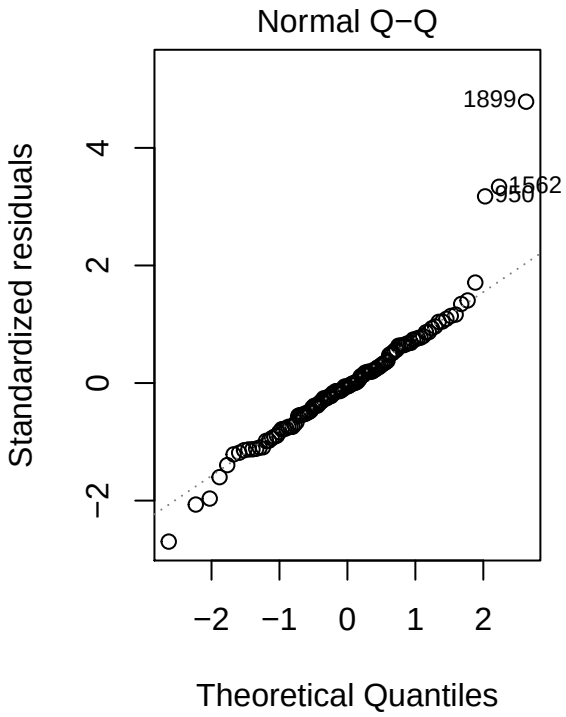



```
data_realistic <- data %>% filter(gestational_age <= 60)
hist(data_realistic$gestational_age, breaks=25)
```




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

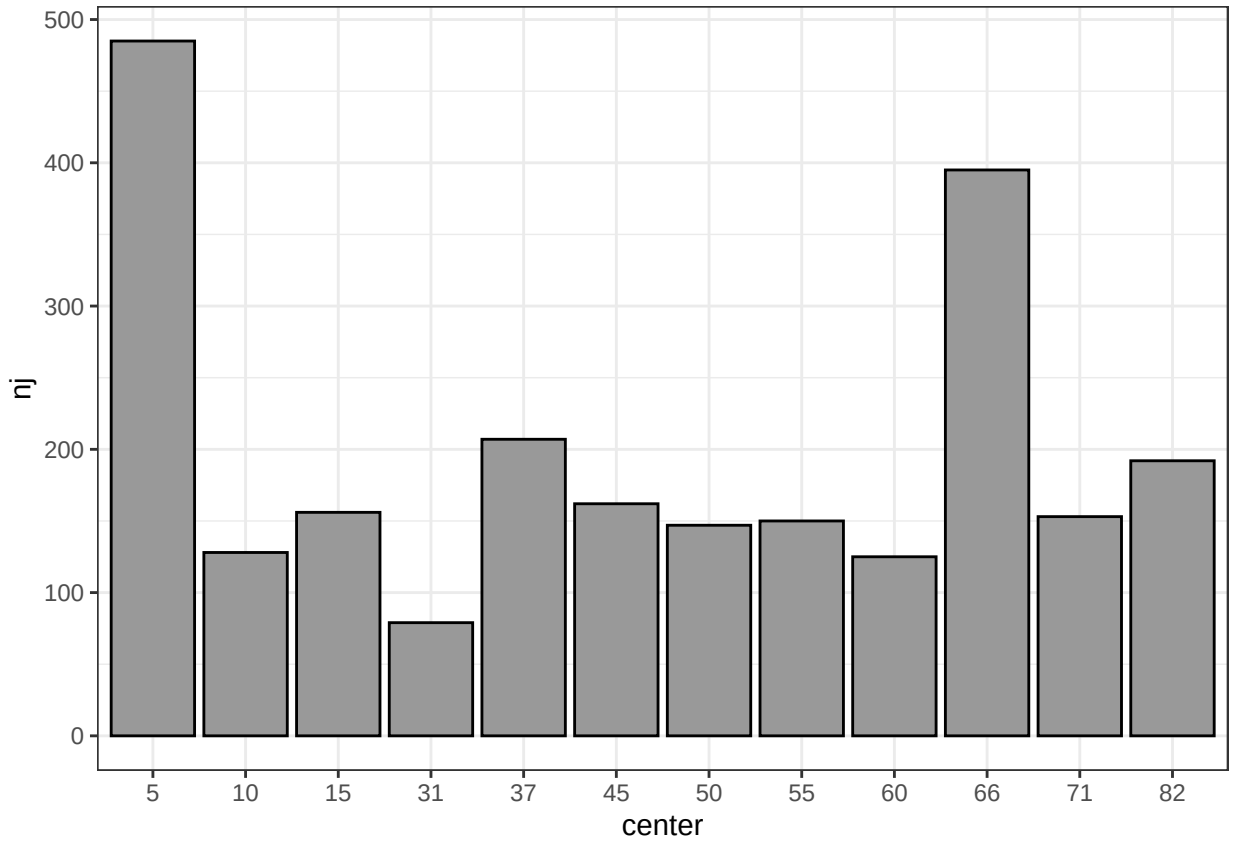
```
## Warning: not plotting observations with leverage one:
##   104, 105
```



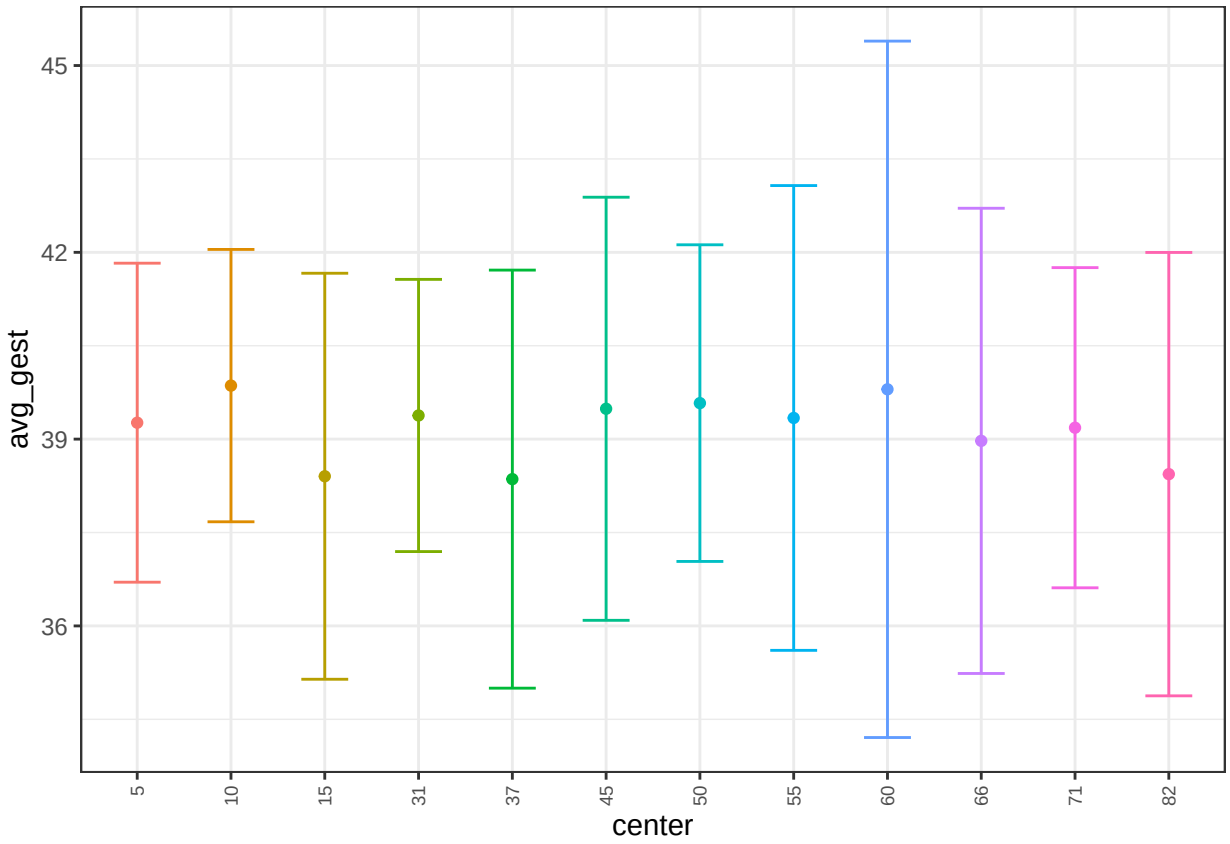

```
# summary stats for outcome by center
samp_stats <- data %>% group_by(center) %>%
  summarise(nj=n(), avg_gest=mean(gestational_age),
            variance=var(gestational_age)) %>% data.frame()

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

```
##    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
```

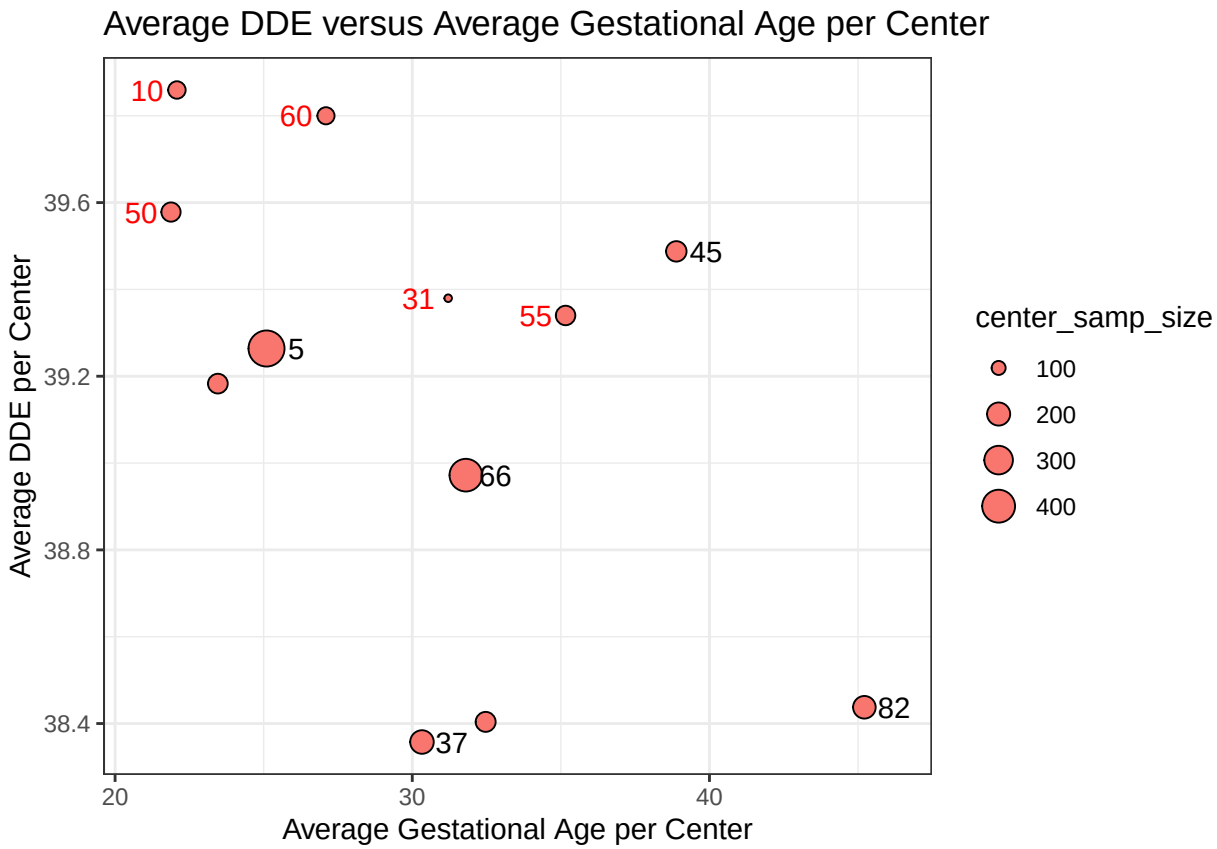
```
samp_stats %>% ggplot(aes(x=center, y=nj)) +
  geom_bar(stat="identity", color="black", fill="#999999") +
  theme_bw()
```



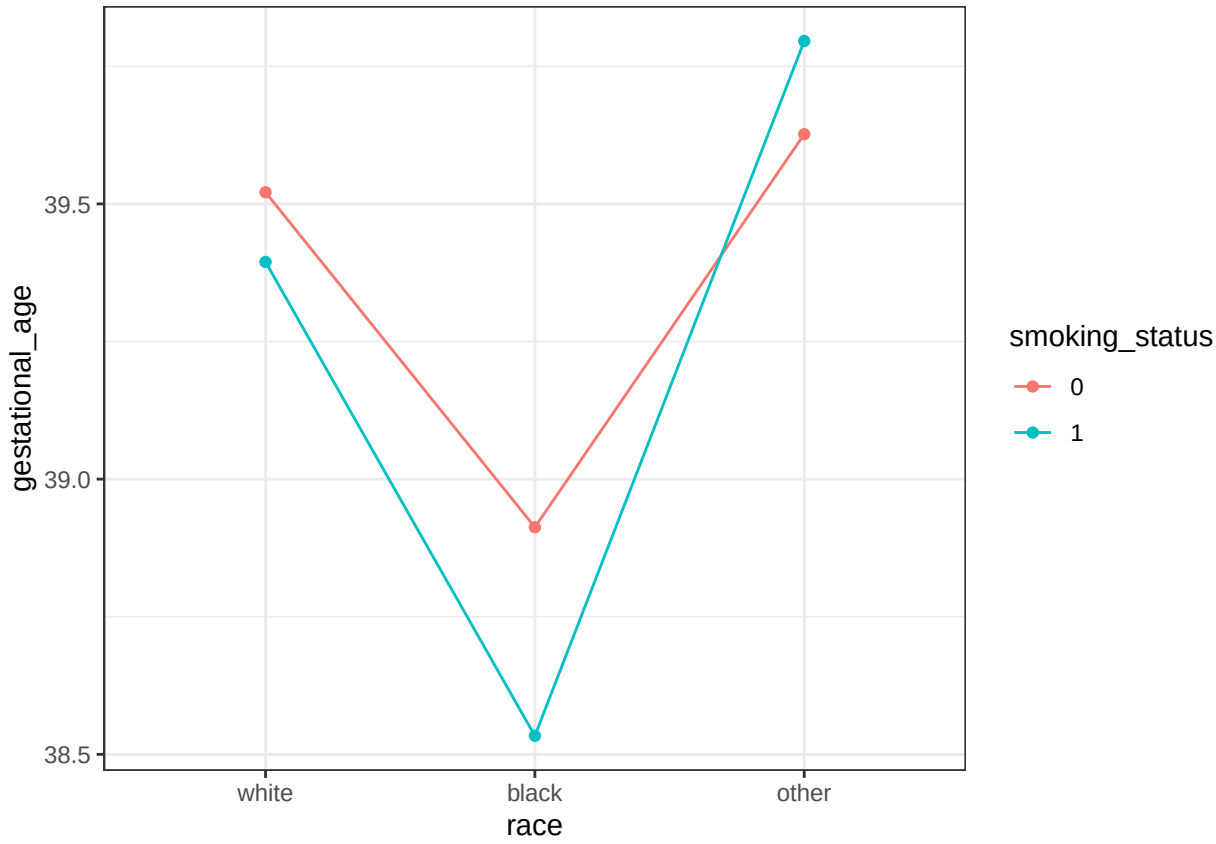

```
# plot summary stats
g1 <- samp_stats %>% mutate(se=sqrt(variance)) %>%
  ggplot(aes(x=center, y=avg_gest, color=center)) +
  geom_point() +
  theme_bw() +
  geom_errorbar(aes(ymin=avg_gest - se, ymax=avg_gest + se), width=.5) +
  theme(legend.position="none",
        axis.text.x=element_text(angle=90,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



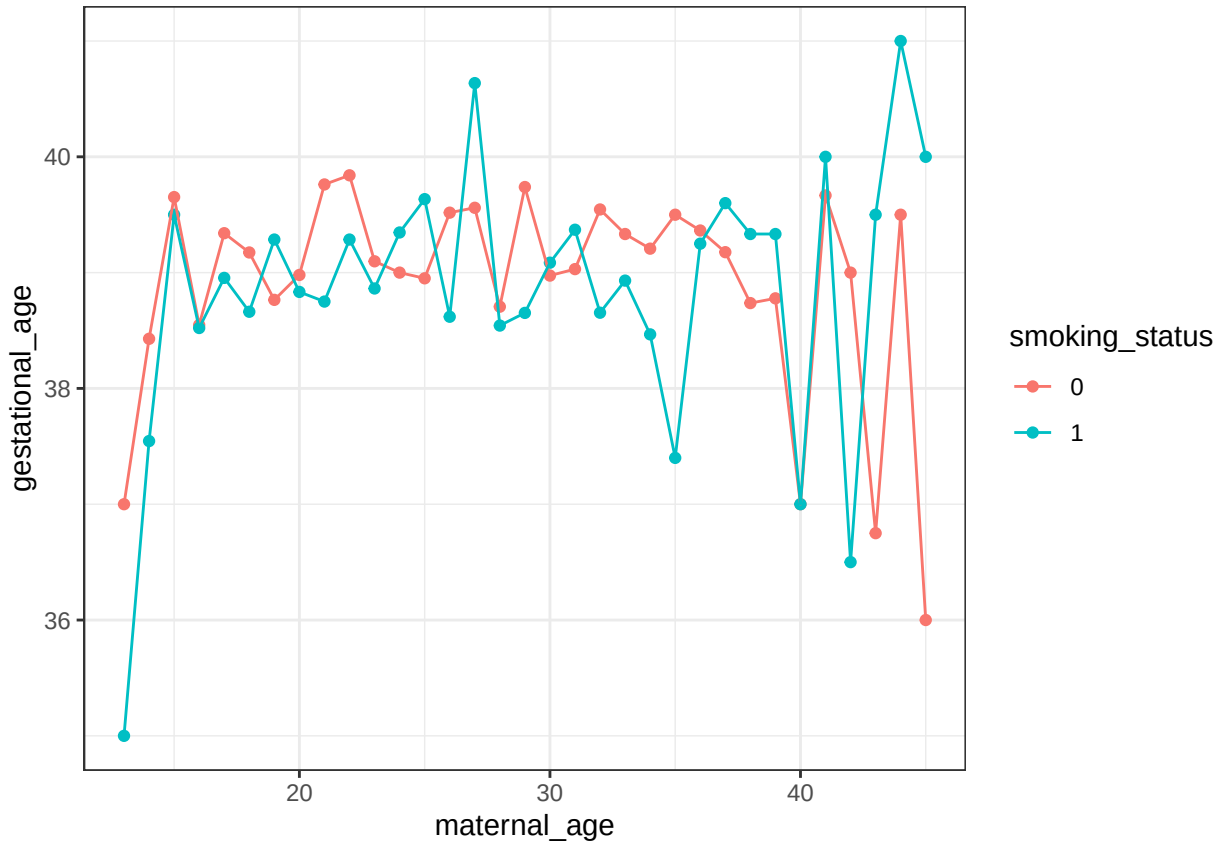

```
# summary stats for outcome and dde by center
samp_stats <- data %>% group_by(center) %>%
  summarise(center_samp_size=n(), avg_gest=mean(gestational_age),
            avg_dde=mean(dde))

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")
```



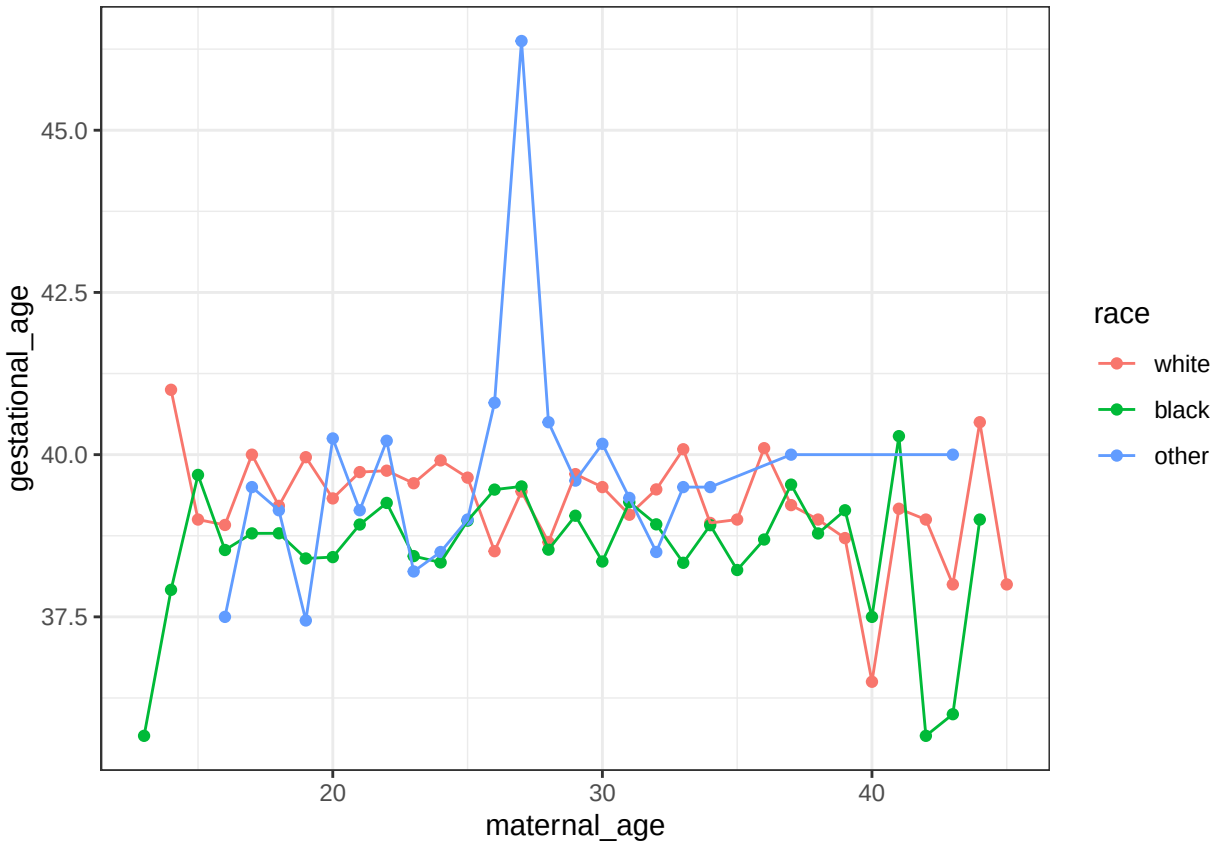

```
# relationship between race and smoking status
# no constant (or parallel), so indicates interactive effect
data %>% ggplot() +
  aes(x=race, color=smoking_status, group=smoking_status, y=gestational_age) +
  stat_summary(fun.y=mean, geom="point") +
  stat_summary(fun.y=mean, geom="line") +
  theme_bw()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




```
# 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()
```



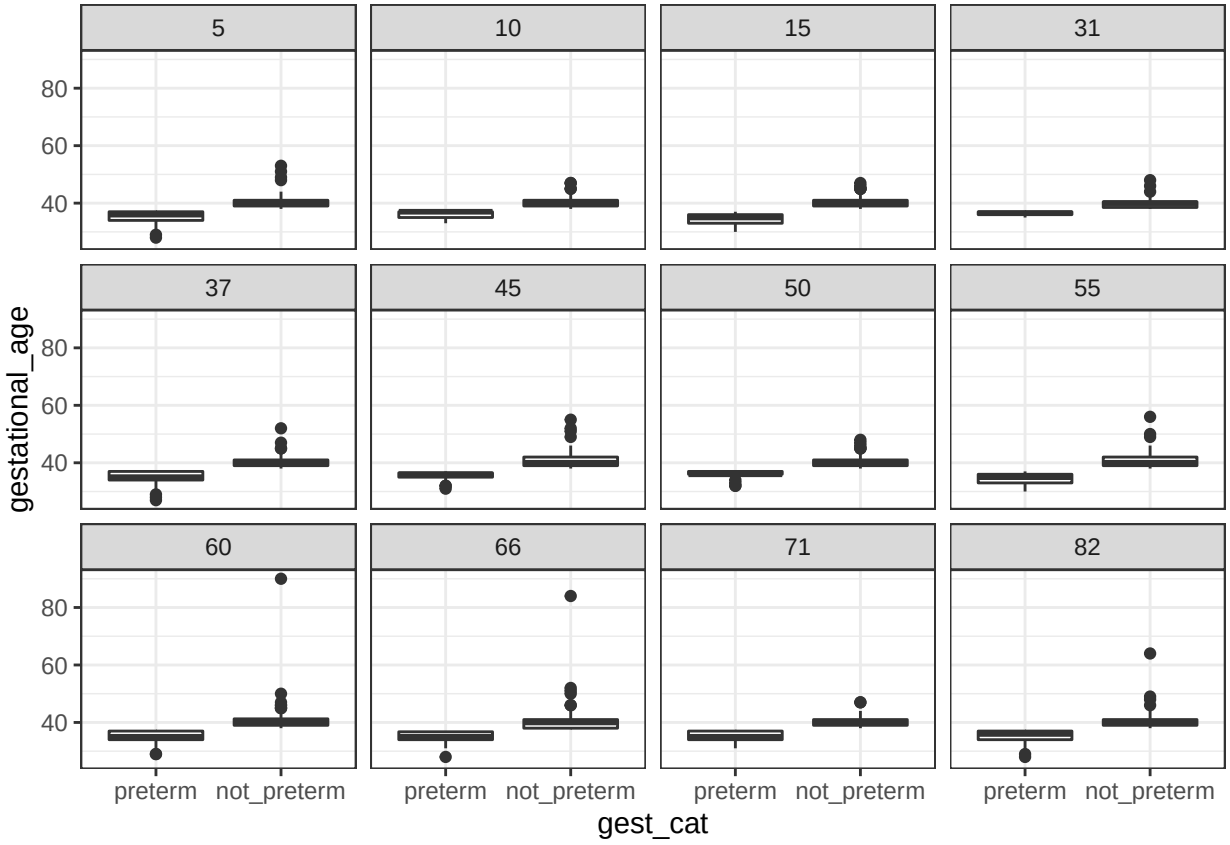

```
# transform gestational age to multi-class variable
# ncbi.nlm.nih.gov/books/NBK279571/ (cutoffs for pre-term pregnancies)
data <- data %>%
  mutate(gest_cat=cut(gestational_age, breaks=c(-Inf, 37, Inf), labels=c("preterm", "not_preterm"))) %>%
  rowwise() %>%
  mutate(min_pcb=min(pcb_028, pcb_052, pcb_074, pcb_105, pcb_118, pcb_153, pcb_170, pcb_180, pcb_194,
                     pcb_203),
         max_pcb=max(pcb_028, pcb_052, pcb_074, pcb_105, pcb_118, pcb_153, pcb_170, pcb_180, pcb_194,
                     pcb_203),
         avg_pcb=mean(c(pcb_028, pcb_052, pcb_074, pcb_105, pcb_118, pcb_153, pcb_170, pcb_180, pcb_194
                        pcb_203)))

data %>% group_by(center, race) %>% summarise(n_cat=n())
```

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

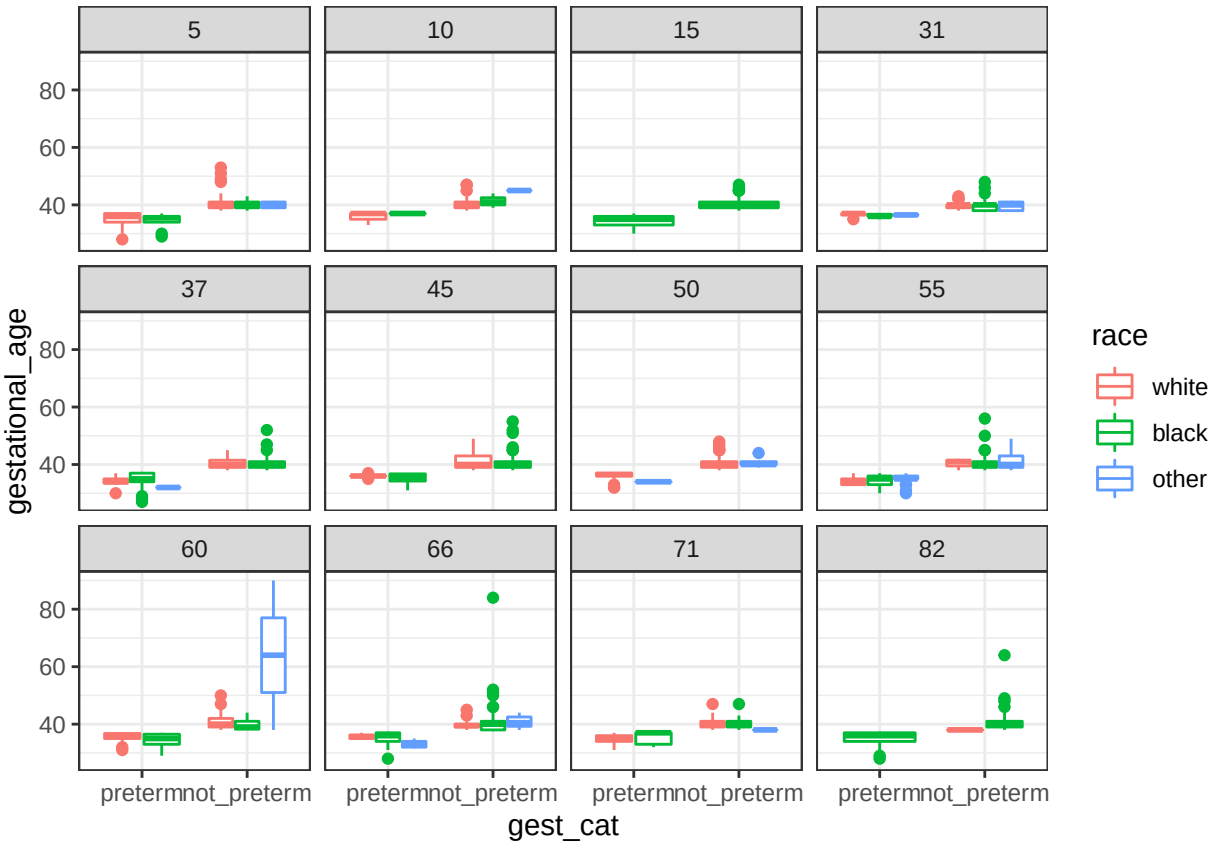
```
## # A tibble: 31 x 3
## # Groups:   center [12]
##    center race  n_cat
##    <fct>  <fct> <int>
##  1 5      white   431
##  2 5      black    50
##  3 5      other     4
##  4 10     white   122
##  5 10     black     5
##  6 10     other     1
##  7 15     black   156
##  8 31     white    21
```

```
##  9 31     black    39
## 10 31     other    19
## # ... with 21 more rows
```

```
data %>% ggplot(aes(x=gest_cat, y=gestational_age)) + geom_boxplot() +
  facet_wrap(. ~ center) +
  theme_bw()
```




```
data %>% ggplot(aes(x=gest_cat, y=gestational_age, color=race)) + geom_boxplot() +
  facet_wrap(. ~ center) +
  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
##            88.0           93.4            0.0            0.0            3.0
##    cholesterol         center            pcb
```

```
##           31.1          100.0           44.1
##
##  Coefficient posterior expected values:
##     (Intercept)              dde    triglycerides        raceblack
##       2.9852235       -0.0092676       -0.0024964        0.0000000
##       raceother     maternal_age  smoking_status1      cholesterol
##       0.0000000        0.0000000       -0.0061287        0.0007817
##        center10         center15         center31         center37
##       1.0416247       -1.1513071        0.3569975       -1.0906632
##        center45         center50         center55         center60
##      -0.4680790        0.0729607       -0.7855366       -0.4834744
##        center66         center71         center82              pcb
##      -0.6084216       -0.1328971       -0.8154005       -0.0446671
```

```
# model w/ random intercept for centers
m1 <- glmer(gest_cat ~ I(triglycerides/100) + I(dde/100) + avg_pcb + race + (1|center), family=binomial
            control=glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)), data=data)
summary(m1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: gest_cat ~ I(triglycerides/100) + I(dde/100) + avg_pcb + race +
##     (1 | center)
##    Data: data
## Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
##      AIC      BIC   logLik deviance df.resid
##   2439.9   2480.3  -1212.9   2425.9     2372
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0675  0.3595  0.4517  0.5652  1.0432
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  center (Intercept) 0.05633  0.2373
## Number of obs: 2379, groups:  center, 12
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.42575    0.21102  11.496  < 2e-16 ***
## I(triglycerides/100) -0.22446    0.06572  -3.416 0.000636 ***
## I(dde/100)           -0.58075    0.26072  -2.227 0.025915 *
## avg_pcb              -0.46225    0.37464  -1.234 0.217253
## raceblack            -0.63287    0.15910  -3.978 6.95e-05 ***
## raceother            -0.42163    0.27136  -1.554 0.120236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) I(t/100) I(d/100) avg_pc rcblck
## I(trgl/100) -0.688
## I(dde/100)  -0.125 -0.095
## avg_pcb     -0.197 -0.130   -0.302
```

```
## raceblack   -0.494  0.261   -0.168   -0.116
## raceother   -0.265  0.040   -0.073    0.089  0.341
```

```
# model w/ random slope for race
m2 <- glmer(gest_cat ~ I(triglycerides/100) + I(dde/100) + avg_pcb + race + (0 + race|center),
            family=binomial, control=glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),
            data=data)
```

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

```
summary(m2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: gest_cat ~ I(triglycerides/100) + I(dde/100) + avg_pcb + race +
##     (0 + race | center)
##    Data: data
## Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
##      AIC      BIC   logLik deviance df.resid
##   2449.3   2518.6  -1212.6   2425.3     2367
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9785  0.3664  0.4479  0.5622  1.0639
##
## Random effects:
##  Groups Name      Variance Std.Dev. Corr
##  center racewhite 0.00000  0.0000
##         raceblack 0.08020  0.2832    NaN
##         raceother 0.08078  0.2842    NaN 1.00
## Number of obs: 2379, groups:  center, 12
##
## Fixed effects:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.49809    0.19529  12.792  < 2e-16 ***
## I(triglycerides/100) -0.23107    0.06621  -3.490 0.000483 ***
## I(dde/100)           -0.59055    0.25993  -2.272 0.023091 *
## avg_pcb              -0.39147    0.36188  -1.082 0.279348
## raceblack            -0.71052    0.16458  -4.317 1.58e-05 ***
## raceother            -0.58819    0.35118  -1.675 0.093952 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) I(t/100) I(d/100) avg_pc rcblck
## I(trgl/100) -0.751
## I(dde/100)  -0.110 -0.095
## avg_pcb     -0.276 -0.130   -0.290
## raceblack   -0.348  0.217   -0.220    0.008
## raceother   -0.217  0.085   -0.070    0.121  0.072
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```r
# model w/ random slope for race and random intercept for centers
m3 <- glmer(gest_cat ~ I(triglycerides/100) + I(dde/100) + avg_pcb + race + (1|center) + (race|center),
            family=binomial, control=glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),
            data=data)
```

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

```r
summary(m3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: gest_cat ~ I(triglycerides/100) + I(dde/100) + avg_pcb + race +
##     (1 | center) + (race | center)
##    Data: data
## Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
##      AIC      BIC   logLik deviance df.resid
##   2451.2   2526.3  -1212.6   2425.2     2366
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0074  0.3665  0.4482  0.5621  1.0620
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  center   (Intercept) 0.01218  0.1103
##  center.1 (Intercept) 0.00000  0.0000
##           raceblack   0.06396  0.2529    NaN
##           raceother   0.05121  0.2263    NaN 1.00
## Number of obs: 2379, groups:  center, 12
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.47780    0.21673  11.433  < 2e-16 ***
## I(triglycerides/100) -0.22921    0.06655  -3.444 0.000573 ***
## I(dde/100)           -0.58484    0.26116  -2.239 0.025129 *
## avg_pcb              -0.42371    0.38585  -1.098 0.272153
## raceblack            -0.68640    0.19435  -3.532 0.000413 ***
## raceother            -0.53830    0.40635  -1.325 0.185265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) I(t/100) I(d/100) avg_pc rcblck
## I(trgl/100) -0.715
## I(dde/100)  -0.134 -0.086
## avg_pcb     -0.103 -0.157   -0.303
## raceblack   -0.501  0.247   -0.132   -0.183
## raceother   -0.378  0.127   -0.018   -0.066  0.324
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```r
# simple model
m4 <- glm(gest_cat ~ 1 + I(triglycerides/100) + I(dde/100) + center + avg_pcb + race, family=binomial,
```

```
          data=data)
summary(m4)
```

```
##
## Call:
## glm(formula = gest_cat ~ 1 + I(triglycerides/100) + I(dde/100) +
##     center + avg_pcb + race, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3130   0.4571   0.6027   0.7371   1.2235
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.63398    0.22419  11.749  < 2e-16 ***
## I(triglycerides/100) -0.22567    0.06619  -3.410 0.000651 ***
## I(dde/100)           -0.51019    0.26462  -1.928 0.053852 .
## center10              0.47109    0.32128   1.466 0.142571
## center15             -0.72929    0.26871  -2.714 0.006647 **
## center31              0.26671    0.36369   0.733 0.463354
## center37             -0.80638    0.22518  -3.581 0.000342 ***
## center45             -0.11957    0.26178  -0.457 0.647844
## center50             -0.25566    0.25978  -0.984 0.325050
## center55             -0.35309    0.30037  -1.176 0.239786
## center60             -0.35671    0.26866  -1.328 0.184266
## center66             -0.20117    0.22323  -0.901 0.367509
## center71             -0.12246    0.25346  -0.483 0.628998
## center82             -0.73297    0.26029  -2.816 0.004863 **
## avg_pcb              -0.70217    0.38226  -1.837 0.066225 .
## raceblack            -0.44182    0.17778  -2.485 0.012947 *
## raceother            -0.35702    0.30805  -1.159 0.246474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2503.5  on 2378  degrees of freedom
## Residual deviance: 2399.9  on 2362  degrees of freedom
## AIC: 2433.9
##
## Number of Fisher Scoring iterations: 4
```

```
# model fits
BIC(m1); BIC(m2); BIC(m3); BIC(m4)
```

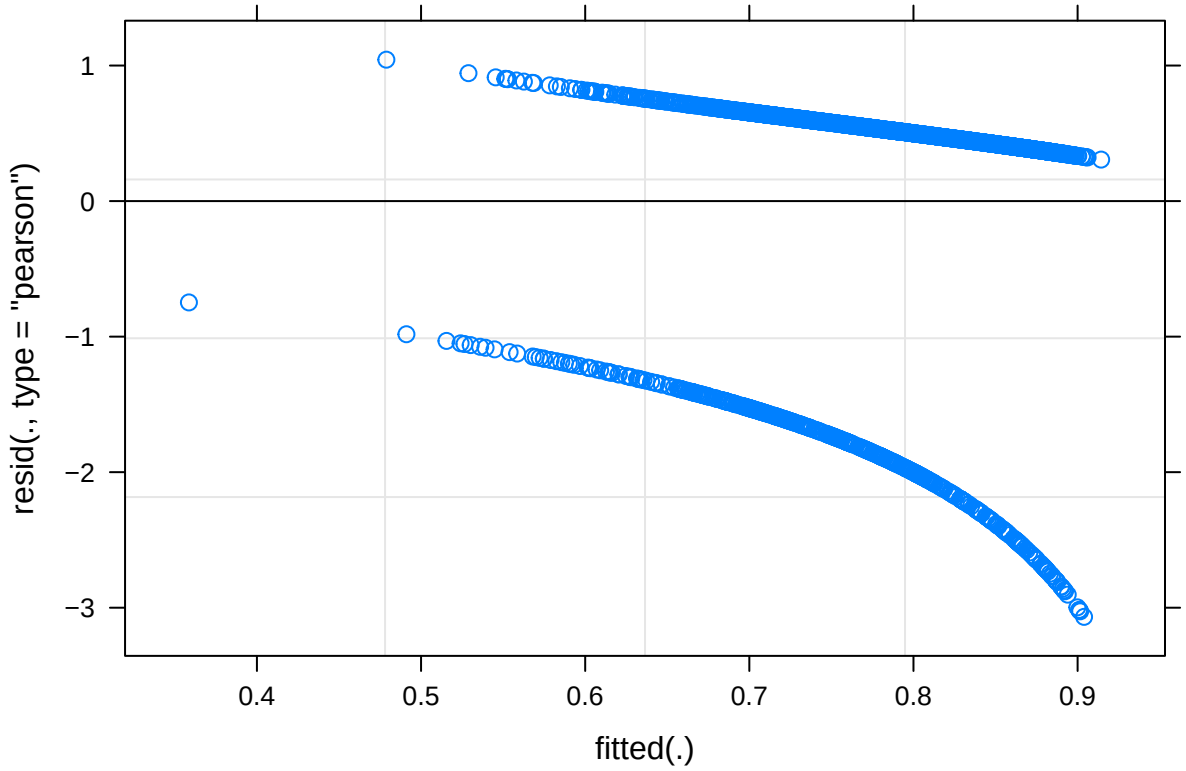
```
## [1] 2480.282
```

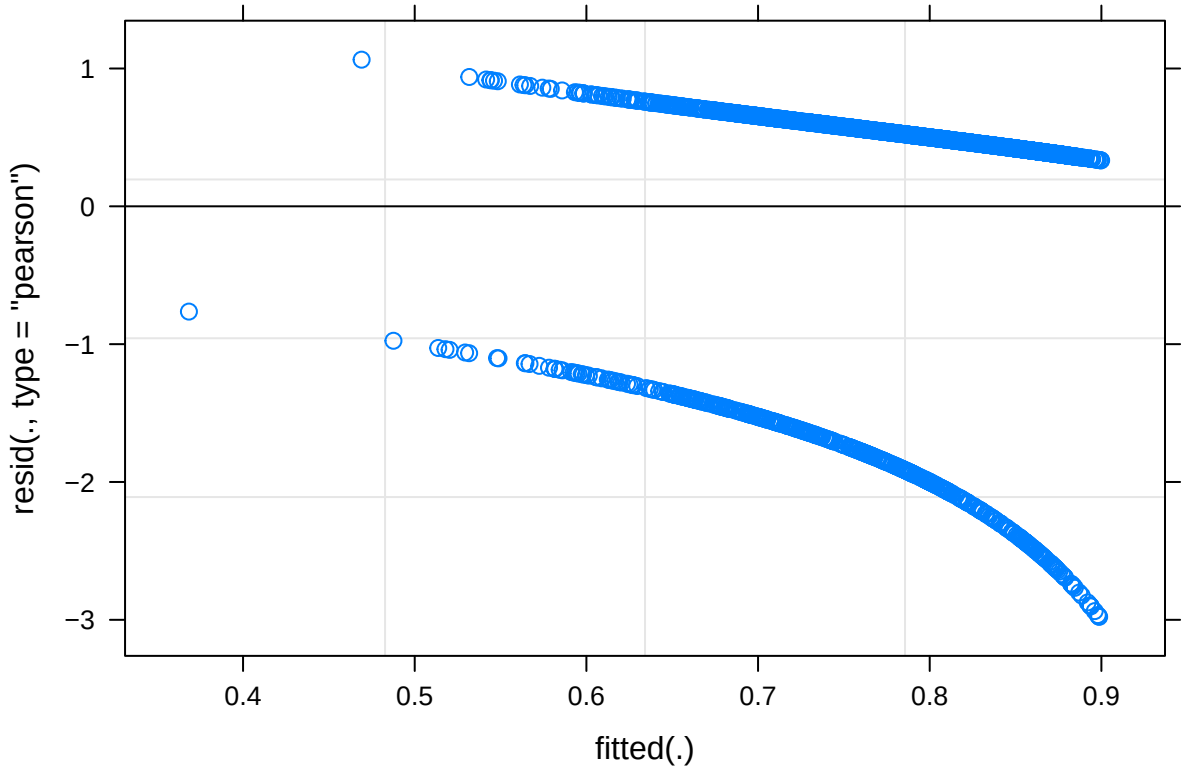
```
## [1] 2518.553
```

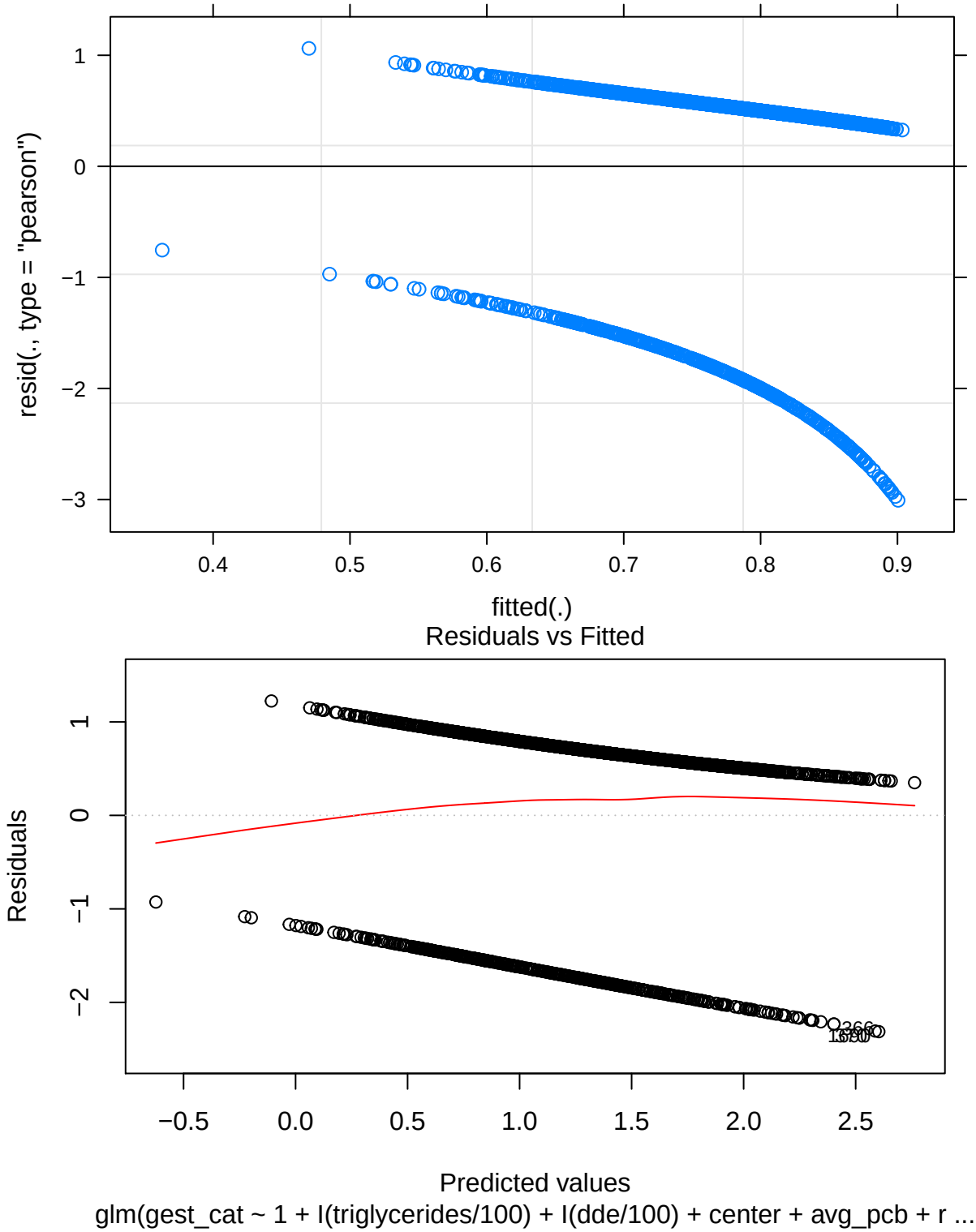
```
## [1] 2526.294
```

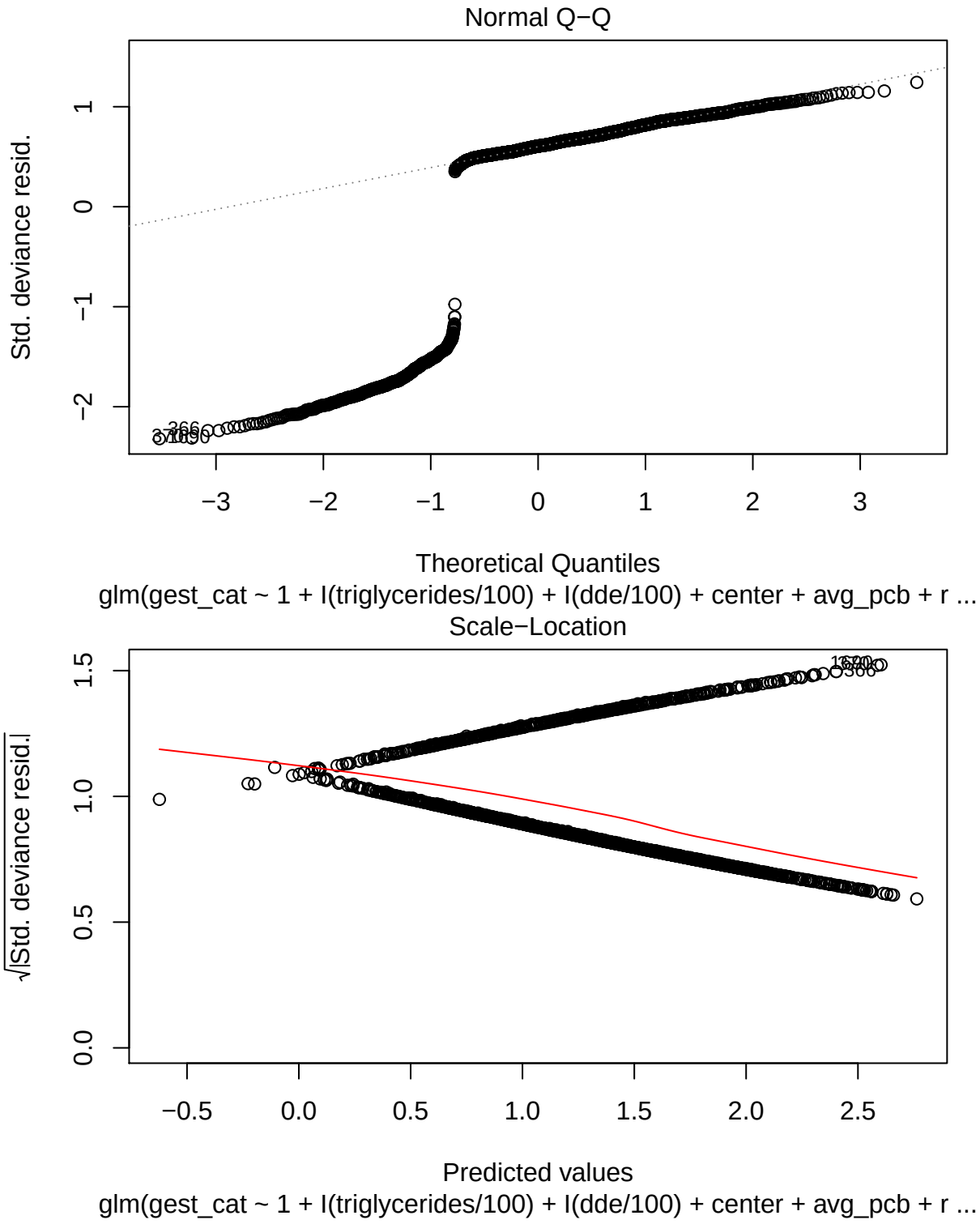
```
## [1] 2532.083
```

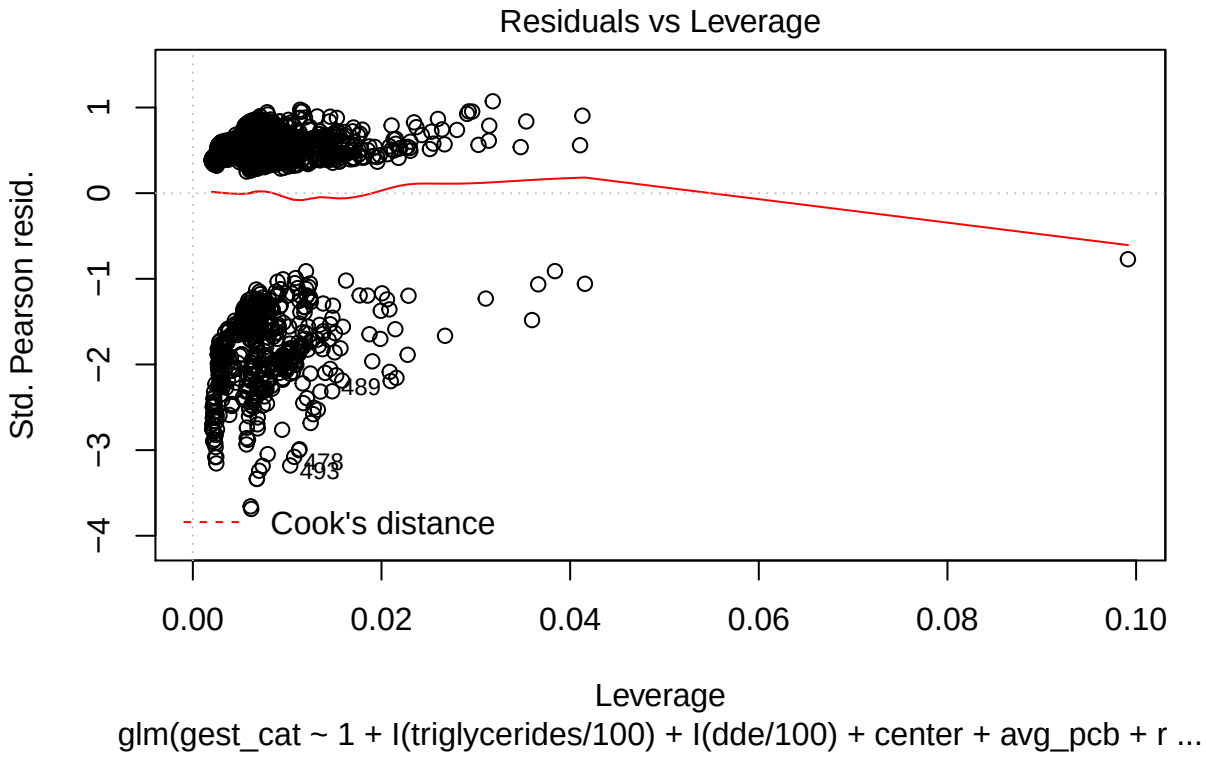
```
# residual plots
plot(m1); plot(m2); plot(m3); plot(m4)
```


resid(., type = "pearson")
1
0
−1
−2
−3
0.4
0.5
0.6
0.7
0.8
0.9
fitted(.)


resid(., type = "pearson")
1
0
−1
−2
−3
0.4
0.5
0.6
0.7
0.8
0.9
fitted(.)


resid(., type = "pearson")
1
0
−1
−2
−3
0.4
0.5
0.6
0.7
0.8
0.9
fitted(.)
Residuals vs Fitted
Residuals
1
0
−1
−2
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Predicted values
glm(gest_cat ~ 1 + I(triglycerides/100) + I(dde/100) + center + avg_pcb + r ...

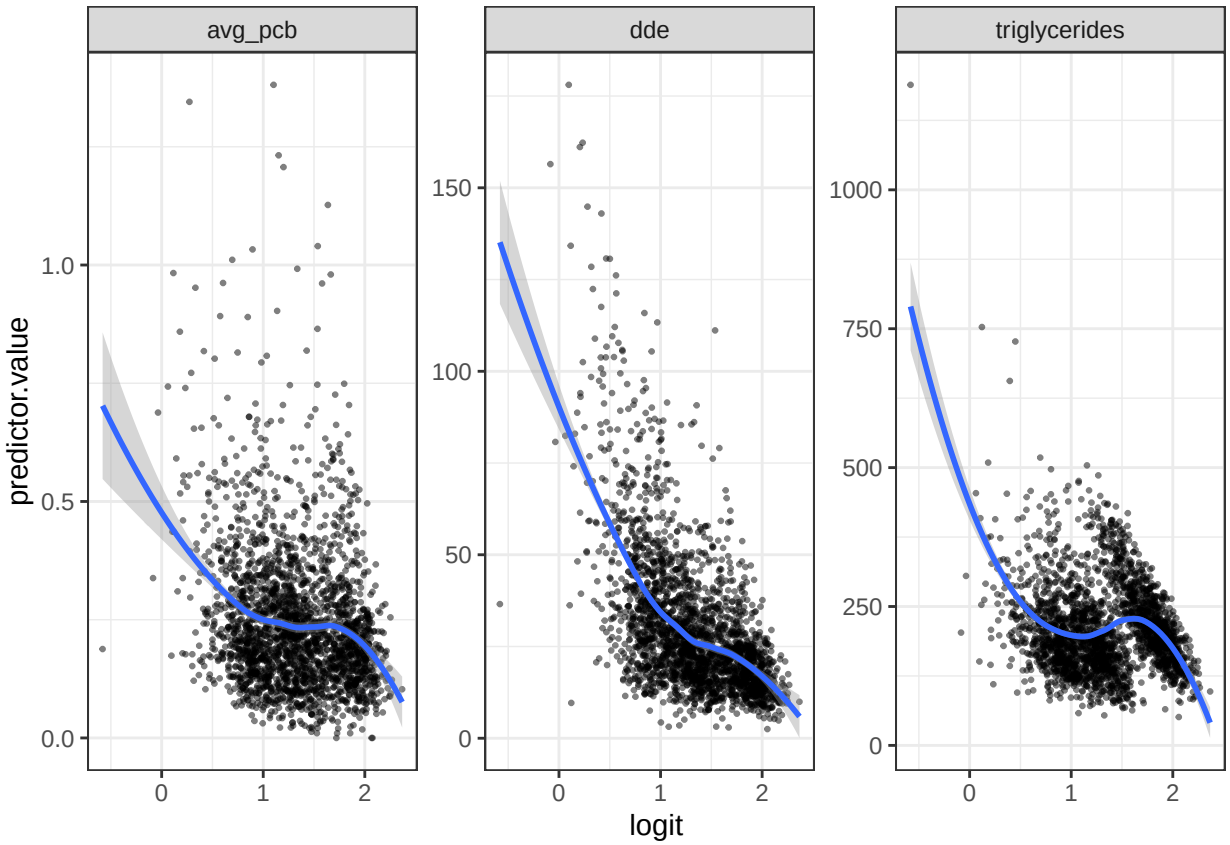

Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(gest_cat ~ 1 + I(triglycerides/100) + I(dde/100) + center + avg_pcb + r ...
Scale−Location
√|Std. deviance resid.|
Predicted values
glm(gest_cat ~ 1 + I(triglycerides/100) + I(dde/100) + center + avg_pcb + r ...



##Checking Logistic Model Assumptions

Linearity of Covariates to Logit:

```
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) 1.095953  1        1.046878
## I(dde/100)           1.164575  1        1.079155
## avg_pcb              1.176784  1        1.084797
## race                 1.144489  2        1.034315
```