ReMAPP Statistical Analysis Plan R Code - Aim 1

Table of Contents

[1 Data preparation 1](#_Toc91599643)

[1.1 Data description 1](#_Toc91599644)

[1.2 Data cleaning 2](#_Toc91599645)

[1.3 Simulation data 2](#_Toc91599646)

[2 Test CBC-hemoglobin heterogeneity across sites 2](#_Toc91599647)

[2.1 Data preparation 2](#_Toc91599648)

[2.2 Variance component analysis on linear mixed-effects model 2](#_Toc91599649)

[3 Calibrate POC-hemoglobin to CBC-hemoglobin 4](#_Toc91599650)

[3.1 Data preparation 4](#_Toc91599651)

[3.2 Visual inspection 4](#_Toc91599652)

[3.3 Wilcoxon signed rank test 6](#_Toc91599653)

[4 Centile plots of hemoglobin against gestational weeks 7](#_Toc91599654)

[4.1 Data preparation 7](#_Toc91599655)

[4.2 Fractional polynomial regression 7](#_Toc91599656)

[4.3 Estimate the centiles at given gestational weeks 10](#_Toc91599657)

[4.4 Diagnostic plot 11](#_Toc91599658)

[4.5 Sensitivity analysis 13](#_Toc91599659)

[4.5.1 Excluding one site at a time 13](#_Toc91599660)

[4.5.2 Excluding subjects with preterm delivery 15](#_Toc91599661)

# 1 Data preparation

## 1.1 Data description

Each participant/subject in the analysis data has to satisfy the following conditions:

1. Enrolled into Aim 2 prior to 14 gestational weeks;
2. Has at least one non-missing hemoglobin data point;
3. Meets the 31 baseline healthy criteria;
4. During the pregnancy and 6 weeks postpartum periods, there is NO multiple pregnancy, cancer, HIV, TB, malaria, or severe pregnancy-related complications requiring hospital admission diagnosed.

The final analysis data in long-format is expected to have around 4\*600\*5 = 12,000 observations (4 sites, 600 people per site, 5 visits = 4 ANC visit + 1 PNC visit) in total.

## 1.2 Data cleaning

Remove rows that have gestational weeks less than 10 or greater than 50 (42-week pregnancy + 8-week postpartum).

Remove rows that have hemoglobin less than 5 or greater than 18.

Remove rows that have missing values on gestational weeks.

## 1.3 Simulation data

The simulation data in this demo is based on the limited Ghana site data. After data cleaning, the simulation data set df\_sim used has 1079 rows and 6 columns: subjid (subject id), hb\_cbc (hemoglobin measured in complete blood count), hb\_poc (hemoglobin measured in point-of-care), ga\_wks (gestations weeks), trimes (trimesters), and site. Though we only have one site data, the column site randomly assigns the subject into one of four hypothesized sites for illustration purpose.

**Table** **1**: simulation data example (empty indicates missing)

| subjid | hb\_cbc | hb\_poc | ga\_wks | trimes | site |
| --- | --- | --- | --- | --- | --- |
| GHA0005 |  | 10.07 | 32 | 3 | 4 |
| GHA0006 |  | 12.40 | 30 | 3 | 4 |
| GHA0006 |  | 12.00 | 36 | 3 | 4 |
| GHA0011 |  | 11.10 | 36 | 3 | 2 |
| GHA0012 |  | 9.50 | 25 | 2 | 2 |

# 2 Test CBC-hemoglobin heterogeneity across sites

## 2.1 Data preparation

Question: do we require each subject has equal number of total visits?

## 2.2 Variance component analysis on linear mixed-effects model

Heterogeneity in CBC-hemoglobin across sites is assessed by comparing the percentages of the variance due to inter-subject and inter-site differences estimated by analysis-of-variance techniques. Variance component analysis is based on a liner mixed-effect model, with gestational age treated as fixed effect, and sites and individuals treated as random effects.

# call package lme4 to fit linear mixed effects model  
library(lme4)  
  
# fit linear mixed effects model  
# use site and subject as random effects  
lme\_mod <- lmer(hb\_cbc ~ ga\_wks + (1 | site) + (1 | site:subjid), data = df\_sim, REML = TRUE)  
  
# obtain variance components  
hb\_var <- VarCorr(lme\_mod) %>% as.data.frame()  
# calculate the variance proportion  
hb\_var\_prop <- data.frame(  
 "Variance compoent" = c("Var(subject within site)", "Var(site)", "Var(error)"),  
 Estimates = hb\_var$vcov,  
 "Standard error" = hb\_var$sdcor,  
 check.names = FALSE  
) %>%  
 mutate("Proportion(%)" = Estimates / sum(Estimates) \* 100) %>%  
 mutate(across(c(2:4), ~ round(.x, 3)))  
  
# print table  
tb\_flextable(hb\_var\_prop, "variance components analysis",   
 seq\_id = "tb", bkm = "tb2")

**Table** **2**: variance components analysis

| Variance compoent | Estimates | Standard error | Proportion(%) |
| --- | --- | --- | --- |
| Var(subject within site) | 0.569 | 0.754 | 34.167 |
| Var(site) | 0.005 | 0.067 | 0.272 |
| Var(error) | 1.092 | 1.045 | 65.560 |

Note from the last column of table 2 that the percentage of variance due to inter-site (0.272%) is less than that of inter-individual (34.167%), therefore, we conclude that the heterogeneity of CBC-hemoglobin across sites can be ignored and we can pool the data together. (\*\*\*The simulation data comes from only one Ghana site, so it’s expected that the inter-site variance is very small, this may not be the case when real data is used in the future, but the methods still apply.)

# 3 Calibrate POC-hemoglobin to CBC-hemoglobin

Since it’s concluded from section 2 that the CBC-hemoglobin measurements are homogeneous across sites, so we examine the difference between POC-hemoglobin and CBC-hemoglobin stratified by trimesters (three trimesters and 6-week postpartum), not by sites. (\*\*\*The simulation data has no 6-week postpartum hemoglobin measurements, but the methods still apply once the measurements are included in the real data in the future.)

## 3.1 Data preparation

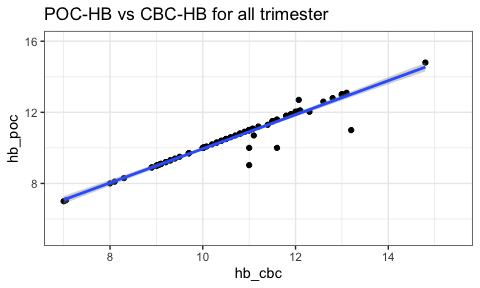
Keep subject visits that have both non-missing CBC-hemoglobin and POC-hemoglobin recorded at the same time points.

# convert data to long-format of hemoglobin  
df\_hb\_calib <- df\_sim %>%  
 filter(!is.na(hb\_poc) & !is.na(hb\_cbc)) %>%  
 select(hb\_cbc, hb\_poc, ga\_wks, trimes) %>%  
 pivot\_longer(c(hb\_cbc, hb\_poc),  
 names\_to = "hb\_type",  
 values\_to = "hb\_value"  
 )

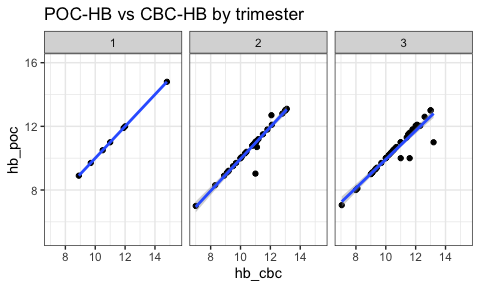
## 3.2 Visual inspection

We first visually inspect if POC-hemoglobin measurements are consistent with the CBC-hemoglobin measurements using point plots and simple linear regressions. It can be observed from the plots that there seems no significant difference between POC-hemoglobin and CBC-hemoglobin, regardless of being stratified by trimesters or not.

# plot of POC hemoglobin vs CBC hemoglobin for all trimesters  
df\_sim %>%  
 ggplot(aes(x = hb\_cbc, y = hb\_poc)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", formula = y ~ x) +  
 ggtitle("POC-HB vs CBC-HB for all trimester") +  
 theme\_bw()



# plot of POC hemoglobin vs CBC hemoglobin by trimester  
df\_sim %>%  
 ggplot(aes(x = hb\_cbc, y = hb\_poc)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", formula = y ~ x) +  
 facet\_wrap(vars(trimes)) +  
 ggtitle("POC-HB vs CBC-HB by trimester") +  
 theme\_bw()



## 3.3 Wilcoxon signed rank test

Next we use *Wilcoxon signed rank test* to test if there is statistically significant difference between the distribution of the POC-hemoglobin data and that of the CBC-hemoglobin data.

# wilcoxon test on all data  
df\_hb\_test\_all <- df\_hb\_calib %>%  
 wilcox.test(hb\_value ~ hb\_type,  
 data = ., alternative = "two.sided", mu = 0,  
 paired = TRUE, exact = FALSE, correct = TRUE, na.rm = TRUE  
 )  
  
# wilcoxon test by trimester  
df\_hb\_test\_tri <- df\_hb\_calib %>%  
 group\_by(trimes) %>%  
 summarise(p.value = wilcox.test(hb\_value ~ hb\_type,  
 alternative = "two.sided", mu = 0,  
 paired = TRUE, exact = FALSE, correct = TRUE,  
 na.rm = TRUE  
 )$p.value)  
  
# combine test results  
tb\_test <- df\_hb\_test\_tri %>%  
 bind\_rows(data.frame(trimes = "All", p.value = df\_hb\_test\_all$p.value)) %>%  
 mutate(p.value = round(p.value, 3))  
  
# print table  
tb\_flextable(tb\_test, "Wilcoxon test result", seq\_id = "tb", bkm = "tb3")

**Table** **3**: Wilcoxon test result

| trimes | p.value |
| --- | --- |
| 1 |  |
| 2 | 0.855 |
| 3 | 0.059 |
| All | 0.067 |

Table 3 shows that the P-values for all tests (empty cell is due to there is only POC-hemoglobin data available in the first trimester), either test by trimester or test for all, are greater than 0.05, therefore we conclude that there is no need to calibrate the POC-hemoglobin measurements at current stage. In the future, if significant difference is detected, then we use linear regression to adjust POC-hemoglobin to the CBC-hemoglobin.

# 4 Centile plots of hemoglobin against gestational weeks

Based on the conclusions from section 2 and 3, we pool the CBC-hemoglobin and calibrated POC-hemoglobin from all sites together, and estimate the 2.5th, 5th, 10th, 50th, 95th and 97.5th centiles of hemoglobin at given gestational weeks during pregnancy and 6-week postpartum.

## 4.1 Data preparation

Data includes three columns: non-missing hemoglobin (CBC + calibrated POC), non-missing gestational weeks, and site.

# convert data to long-format of hemoglobin  
df\_fpr <- df\_sim %>%  
 dplyr::select(hb\_cbc, hb\_poc, ga\_wks, site) %>%  
 pivot\_longer(c(hb\_cbc, hb\_poc),  
 names\_to = "hb\_type",  
 values\_to = "hb\_value"  
 ) %>%  
 dplyr::select(-hb\_type) %>%  
 filter(!is.na(hb\_value), !is.na(ga\_wks))

## 4.2 Fractional polynomial regression

By assuming that the hemoglobin follows a normal distribution at given gestational week , the fractional polynomial regression (FPR) model fits regression models for both the mean parameter and standard deviation parameter . Then estimate the centiles by using the equation

, where is the expected hemoglobin centile at given gestational week , and is the normal equivalent deviate of size .

# call package gamlss to fit FPR model  
library(gamlss)  
  
# fit a second-order FPR model  
con <- gamlss.control(trace = FALSE)  
fpr\_mod <- gamlss(hb\_value ~ fp(ga\_wks, 2),  
 sigma.fo = ~ fp(ga\_wks, 2),  
 data = df\_fpr, control = con  
)  
  
# get model coefficients for mean and sigma  
tb\_coef <- rbind(  
 getSmo(fpr\_mod, what = "mu")$coef,  
 getSmo(fpr\_mod, what = "sigma")$coef  
) %>%  
 as.data.frame() %>%  
 add\_column(c("mu", "sigma"), .before = 1) %>%  
 rename\_with(~ c("parameter", "intercept", "coefficient 1", "coefficient 2"), c(1:4))  
  
# print table  
tb\_flextable(tb\_coef, "FPR coefficients for mu and sigma", seq\_id = "tb", bkm = "tb4")

**Table** **4**: FPR coefficients for mu and sigma

| parameter | intercept | coefficient 1 | coefficient 2 |
| --- | --- | --- | --- |
| mu | 0.73871398 | -0.3785692 | 0.259850101 |
| sigma | -0.01294095 | 0.2806215 | -0.001715067 |

# get model polynomial powers for mean and sigma  
tb\_power <- rbind(  
 getSmo(fpr\_mod, what = "mu")$power,  
 getSmo(fpr\_mod, what = "sigma")$power  
) %>%  
 as.data.frame() %>%  
 add\_column(c("mu", "sigma"), .before = 1) %>%  
 rename\_with(~ c("parameter", "power 1", "power 2"), c(1:3))  
# print table  
tb\_flextable(tb\_power, "FPR powers for mu and sigma", seq\_id = "tb", bkm = "tb5")

**Table** **5**: FPR powers for mu and sigma

| parameter | power 1 | power 2 |
| --- | --- | --- |
| mu | 2 | 2 |
| sigma | -2 | 3 |

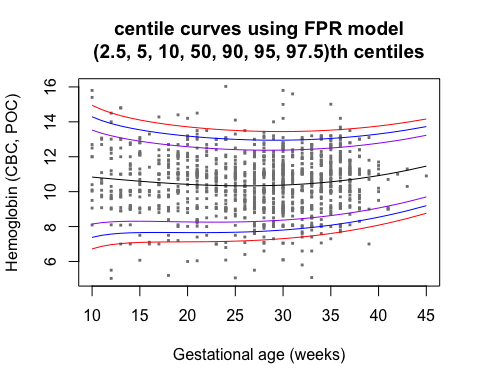
The FPR model for and are

and

respectively.

Given the fitted FPR model, we can plot the centile curves of hemoglobin against the gestational weeks. (\*\*\*Since in the simulation data we have far fewer data points near the boundaries of gestational weeks, we need to be cautious when interpreting the centiles at gestational week around 10 and 50. One alternative is to narrower the range of gestational weeks.)

# make centile plots based on the fitted FPR model  
centiles(fpr\_mod,  
 xvar = df\_fpr$ga\_wks, cent = c(2.5, 5, 10, 50, 90, 95, 97.5),  
 xlab = "Gestational age (weeks)", ylab = "Hemoglobin (CBC, POC)",  
 main = "centile curves using FPR model\n(2.5, 5, 10, 50, 90, 95, 97.5)th centiles",  
 bg = "transparent", legend = FALSE,  
 pch = 15, cex = 0.4, col = gray(0.5),  
 col.cent = c("red", "blue", "purple", "black", "purple", "blue", "red")  
 )

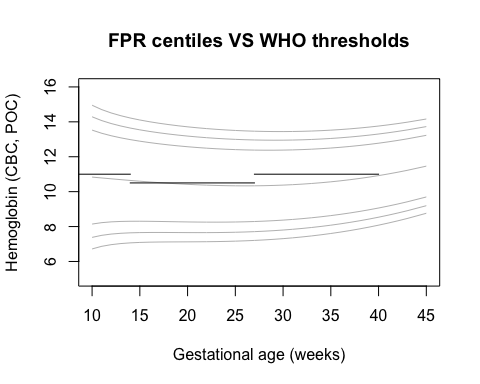


## % of cases below 2.5 centile is 3.173242   
## % of cases below 5 centile is 5.488851   
## % of cases below 10 centile is 10.12007   
## % of cases below 50 centile is 49.14237   
## % of cases below 90 centile is 90.22298   
## % of cases below 95 centile is 95.88336   
## % of cases below 97.5 centile is 98.28473

# compare FPR centiles with WHO thresholds  
centiles(fpr\_mod,  
 xvar = df\_fpr$ga\_wks, cent = c(2.5, 5, 10, 50, 90, 95, 97.5),  
 xlab = "Gestational age (weeks)", ylab = "Hemoglobin (CBC, POC)",  
 main = "FPR centiles VS WHO thresholds",  
 points = FALSE,  
 bg = "transparent", legend = FALSE,  
 col.cent = rep("grey", 7)  
 )

## % of cases below 2.5 centile is 3.173242   
## % of cases below 5 centile is 5.488851   
## % of cases below 10 centile is 10.12007   
## % of cases below 50 centile is 49.14237   
## % of cases below 90 centile is 90.22298   
## % of cases below 95 centile is 95.88336   
## % of cases below 97.5 centile is 98.28473

segments(x0 = 0, y0 = 11, x1 = 14, y1 = 11, col="black")  
segments(x0 = 14, y0 = 10.5, x1 = 27, y1 = 10.5, col="black")  
segments(x0 = 27, y0 = 11, x1 = 40, y1 = 11, col="black")



## 4.3 Estimate the centiles at given gestational weeks

We estimate the 2.5th, 5th, 10th, 50th, 95th and 97.5th centiles at 14, 28, 40 gestational weeks using predictions from the FPR curves.

# centiles at 14, 28, 40 gestational weeks predicted by FPR model  
tb\_cent <- centiles.pred(fpr\_mod,  
 xname = "ga\_wks", xvalues = c(14, 28, 40),  
 type = c("centiles"), cent = c(2.5, 5, 10, 50, 90, 95, 97.5)  
) %>% as.data.frame() %>%   
 round(3) %>%   
 rename\_with(~ c("gestational week", paste0(c(2.5, 5, 10, 50, 90, 95, 97.5), "th")))

## new prediction   
## new prediction

# print table  
tb\_flextable(tb\_cent, "FPR predicted centiles", seq\_id = "tb", bkm = "tb6")

**Table** **6**: FPR predicted centiles

| gestational week | 2.5th | 5th | 10th | 50th | 90th | 95th | 97.5th |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 7.066 | 7.641 | 8.304 | 10.642 | 12.981 | 13.644 | 14.219 |
| 28 | 7.234 | 7.734 | 8.310 | 10.343 | 12.375 | 12.951 | 13.451 |
| 40 | 8.080 | 8.537 | 9.063 | 10.920 | 12.776 | 13.303 | 13.759 |

## 4.4 Diagnostic plot

We use diagnostic plots to examine if the FPR model is a good fit to the hemoglobin centile curves.

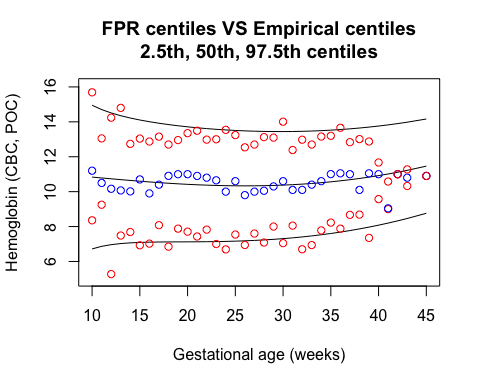
In the “FPR centiles VS Empirical centiles” plot, it can be observed that the FPR curves are less consistent with empirical curves at 2.5th and 97.5th centiles than at 50th centile, especially when gestational weeks are at boundaries near 10 or 50. This might be improved if more data are included or the range of gestational weeks are narrowed.

df\_empi\_cent <- df\_fpr %>%   
 group\_by(ga\_wks) %>%   
 summarise(cent\_2.5 = quantile(hb\_value, probs = c(0.025), na.rm = TRUE),  
 cent\_50 = quantile(hb\_value, probs = c(0.50), na.rm = TRUE),  
 cent\_97.5 = quantile(hb\_value, probs = c(0.975), na.rm = TRUE)) %>%   
 ungroup()

centiles(fpr\_mod,  
 xvar = df\_fpr$ga\_wks, cent = c(2.5, 50, 97.5),  
 xlab = "Gestational age (weeks)", ylab = "Hemoglobin (CBC, POC)",  
 main = "FPR centiles VS Empirical centiles\n2.5th, 50th, 97.5th centiles",  
 points = FALSE,  
 bg = "transparent", legend = FALSE,  
 col.cent = rep("black", 7)  
 )

## % of cases below 2.5 centile is 3.173242   
## % of cases below 50 centile is 49.14237   
## % of cases below 97.5 centile is 98.28473

points(x = df\_empi\_cent$ga\_wks, y = df\_empi\_cent$cent\_2.5, col = "red")  
points(x = df\_empi\_cent$ga\_wks, y = df\_empi\_cent$cent\_50, col = "blue")  
points(x = df\_empi\_cent$ga\_wks, y = df\_empi\_cent$cent\_97.5, col = "red")

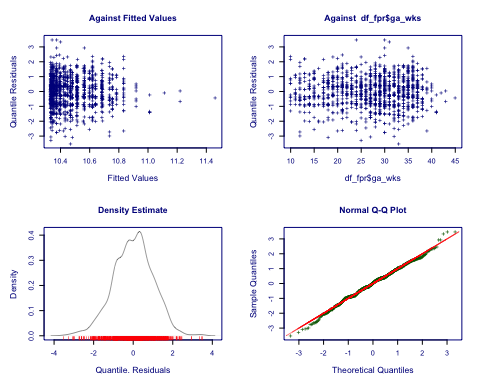


Use plots of normalized quantile residuals to check the adequacy of the fitted fractional polynomial regression model. The four plots are:

* top left: residuals against the fitted values of the mean parameter.
* top right: residuals against the specified covariate.
* bottom left: a kernel density estimate of the residuals.
* bottom right: a QQ-normal plot of the residuals.

The residuals behave well as the top two plots show a random scatter around the horizontal line at 0, and the kernel density estimate of the residuals is approximately normal and the normal QQ-plot is approximately linear. Notice that in the QQ-plot, the points tend to fall off the diagonal line in the extremities, which might be due the existence of extreme values.

newpar <- par(mfrow = c(2, 2), mar = par("mar") + c(0, 1, 0, 0),  
 col.axis = "blue4", col = "blue4", col.main = "blue4",  
 col.lab = "blue4", pch = "+", cex = 0.45, cex.lab = 1.2,  
 cex.axis = 1, cex.main = 1.2)  
plot(fpr\_mod, xvar = df\_fpr$ga\_wks, par = newpar, summaries = FALSE)



par(newpar)

## 4.5 Sensitivity analysis

### 4.5.1 Excluding one site at a time

# sites id  
sites <- as.character(c(1:4))  
# centiles  
cents <- c(2.5, 50, 97.5)  
  
# a function to plot centile curves after excluding certain site  
fpr\_mod\_site <- function(s){  
   
 # remove the selected site data  
 dat <- df\_fpr %>% filter(site != s)  
 # fit a FPR model on the rest data  
 mod <- gamlss(hb\_value ~ fp(ga\_wks, 2),  
 sigma.fo = ~ fp(ga\_wks, 2),  
 data = dat, control = con)  
  
 x <- df\_fpr$ga\_wks  
 # predict hemoglobin centiles by FPR model fitted on all data  
 y\_fpr\_mod <- centiles.pred(fpr\_mod,  
 xname = "ga\_wks", xvalues = x,  
 type = c("centiles"), cent = cents, data = df\_fpr)  
 # predict hemoglobin centiles by FPR model fitted on all data excluding the site  
 y\_mod <- centiles.pred(mod,  
 xname = "ga\_wks", xvalues = x,  
 type = c("centiles"), cent = cents, data = dat)  
  
 # compare the centile curves  
 plot(y\_fpr\_mod[,1], y\_fpr\_mod[,2],  
 ylim = c(6, 16),  
 type="n",  
 xlab = "Gestational week", ylab = "Hemoglobin (CBC+POC)",  
 main = paste0("Excluding site ", s))  
 for(i in 2:(length(cents)+1)) {  
 lines(y\_fpr\_mod[,1], y\_fpr\_mod[,i], col = "black")  
 lines(y\_mod[,1], y\_mod[,i], col = "red")  
 }  
}  
  
  
newpar <- par(mfrow = c(2, 2), mar = par("mar") + c(0, 1, 0, 0),  
 col.axis = "blue4", col = "blue4", col.main = "blue4",  
 col.lab = "blue4", pch = "+", cex = 0.45, cex.lab = 1.2,  
 cex.axis = 1, cex.main = 1.2)  
# remove site 1  
fpr\_mod\_site("1")

## new prediction   
## new prediction   
## new prediction   
## new prediction

# remove site 2  
fpr\_mod\_site("2")

## new prediction   
## new prediction   
## new prediction   
## new prediction

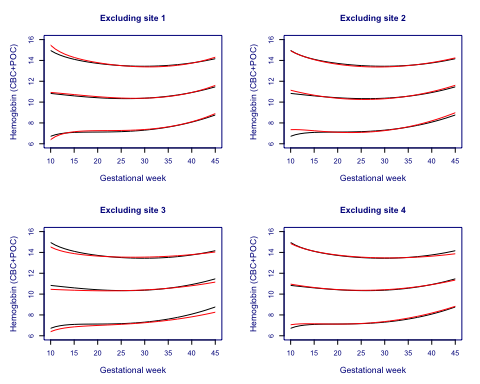
# remove site 3  
fpr\_mod\_site("3")

## new prediction   
## new prediction   
## new prediction   
## new prediction

# remove site 4  
fpr\_mod\_site("4")

## new prediction   
## new prediction   
## new prediction   
## new prediction

par(newpar)



### 4.5.2 Excluding subjects with preterm delivery

The plot currently is not available due to lack of preterm delivery data, but the plotting methods from last section still apply.