Homework #4

Iuliia Oblasova

10/10/2018

```
library(lattice)
library(arm)
```

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

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

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

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

```
## Working directory is /Users/iuliia
```

```
library(pROC)
```

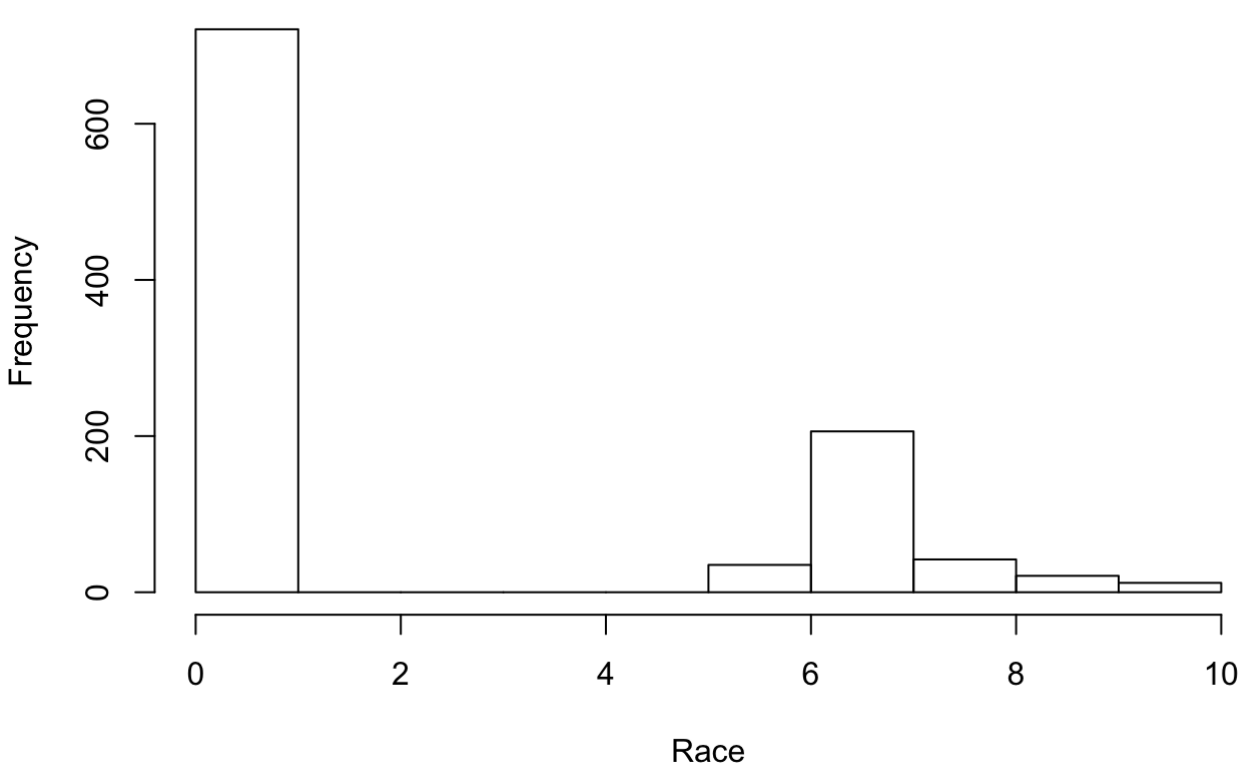
```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(ggplot2)
```

```
smoking <-read.csv("~/Desktop/702 - Data modeling/babiesdata.csv")
smoking$mrace[smoking$mrace==1]<-0
smoking$mrace[smoking$mrace==2]<-0
smoking$mrace[smoking$mrace==3]<-0
smoking$mrace[smoking$mrace==4]<-0
smoking$mrace[smoking$mrace==5]<-0
hist(smoking$mrace, main = "Participating races", xlab="Race")
```



It can be observed that the proportion of participating races is very unbalanced. Exploratory variable - Premature (1 if the pre-term birth given, 0 otherwise). We'll exclude gestation varible from analysis since the outcome variable Prepature has a functional relation with gestational age. Variable "id" and "date" are also excluded since we have a reasonable assumption that these variables do not affect the pre-term birth. All variables with information about fathers' data are also excluded to avoid missing data, as well as variables "data" and "number" since we compare only "smokers" to "non-smokers".

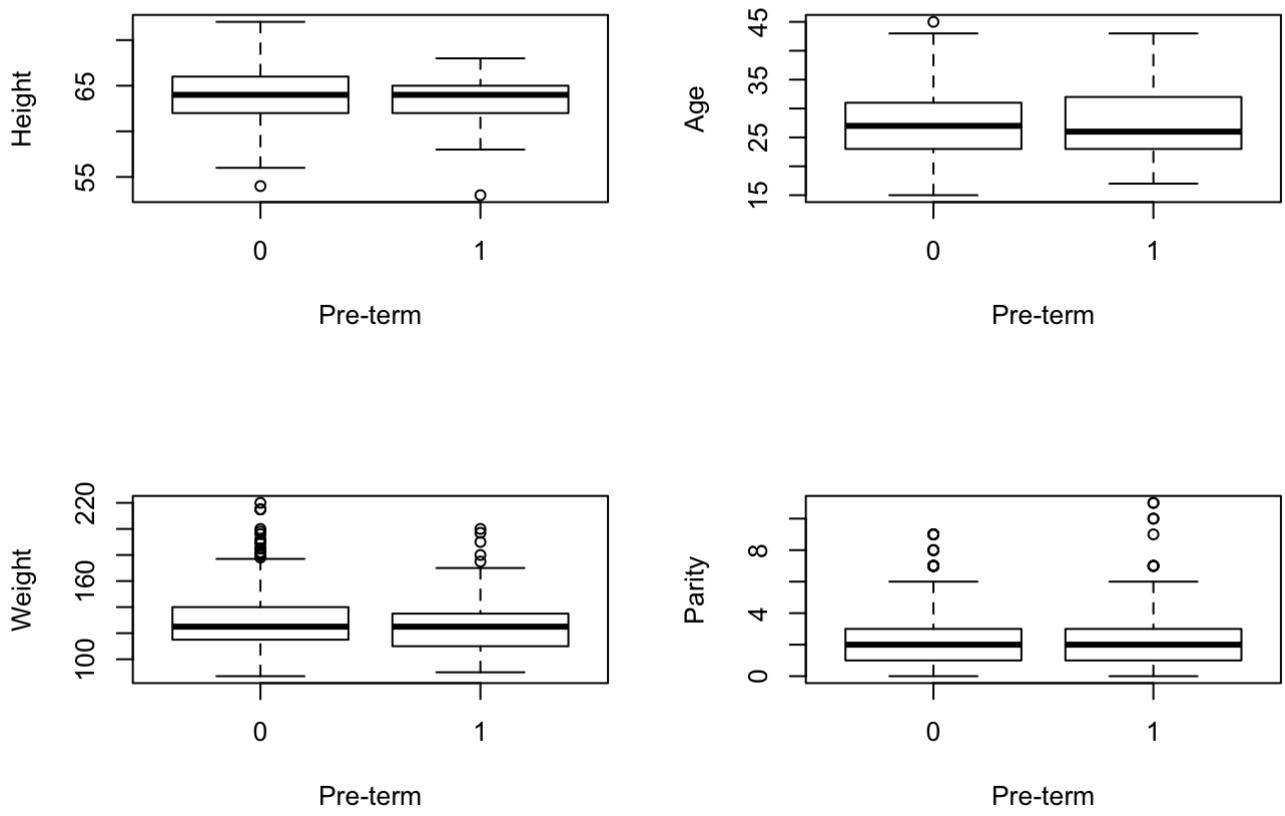
```
smoking[ ,c('gestation','id','date','drace','dage','ded','dht','dwt','bwt.oz','number',
'marital','time')] <- list(NULL)
smoking<-smoking[complete.cases(smoking), ]
dim(smoking)
```

```
## [1] 880   9
```

The resulting table contains 880 of 9 variables whereas the original data set had 1038 observation of 21 variable. Therefore, our model was significantly simplified for the purposes of the analysis.

It should be noted that the number of given preterm births is three times less than the number of births given on time.

```
par(mfcol=c(2,2))
boxplot(mht~Premature, data=smoking, xlab="Pre-term", ylab="Height")
boxplot(mpregwt~Premature, data=smoking, xlab="Pre-term", ylab="Weight")
boxplot(mage~Premature, data=smoking, xlab="Pre-term", ylab="Age")
boxplot(parity~Premature, data=smoking, xlab="Pre-term", ylab="Parity")
```

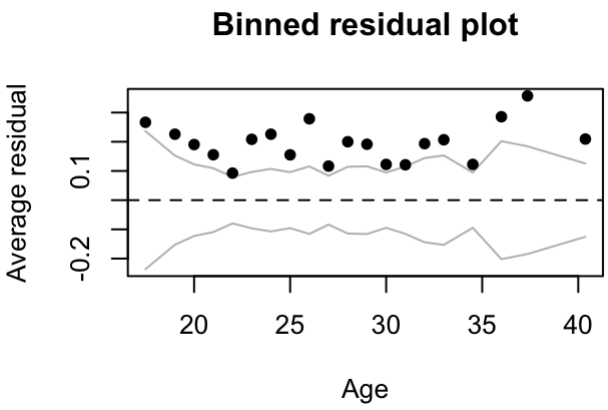


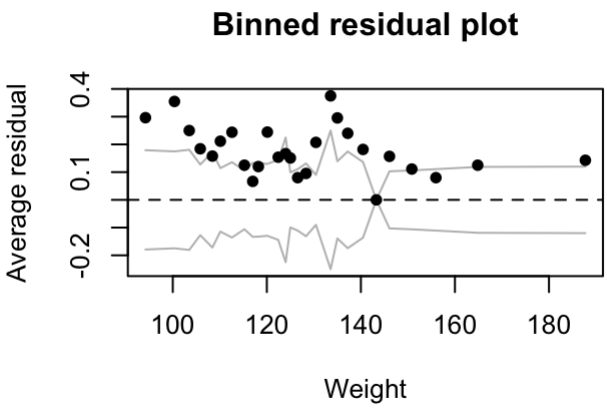
No differences in medians is observed. Center continious predictors mage, mht, mpregwt and parity.

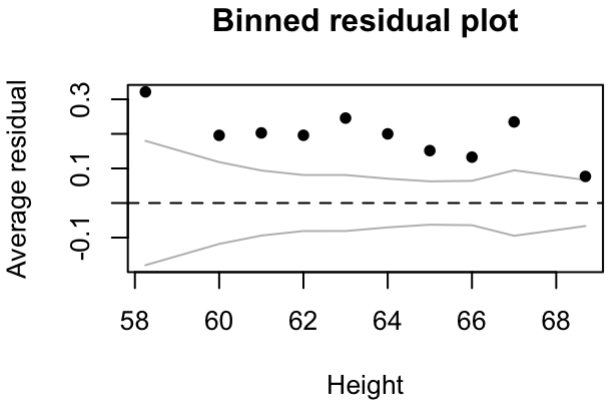
```
smoking$mage.c = smoking$mage-mean(smoking$mage)
smoking$mht.c = smoking$mht-mean(smoking$mht)
smoking$mpregwt.c = smoking$mpregwt-mean(smoking$mpregwt)
smoking$parity.c = smoking$parity-mean(smoking$parity)
```

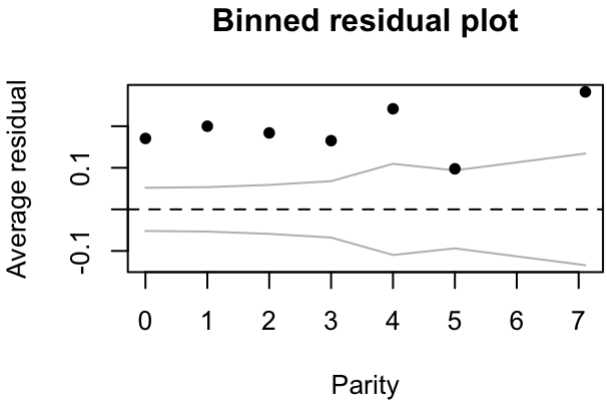
Let's make a binned plots for continious variables.

```
par(mfcol=c(2,2))
binnedplot(x=smoking$mage,y=smoking$Premature,xlab="Age")
binnedplot(x=smoking$mht,y=smoking$Premature,xlab="Height")
binnedplot(x=smoking$mpregwt,y=smoking$Premature,xlab="Weight")
binnedplot(x=smoking$parity,y=smoking$Premature,xlab="Parity")
```









No obvious transformation suggested, so let's plot the first model. First, let's define the levels of the categorical variables.

```
smoking$smoke = factor(smoking$smoke, levels = c("1","0"))
smoking$mrace = factor(smoking$mrace, levels = c("0","6","7","8","9"))
smoking$med = factor(smoking$med, levels = c("0","1","2","3","4","5","6","7","8","9"))
smoking$inc = factor(smoking$inc, levels = c("0","1","2","3","4","5","6","7","8","9"))
```

```
preterm1 = glm(Premature ~ mage.c + mht.c + mpregwt.c + parity.c + mrace + med + inc, da
ta = smoking, family = binomial)
summary(preterm1)
```

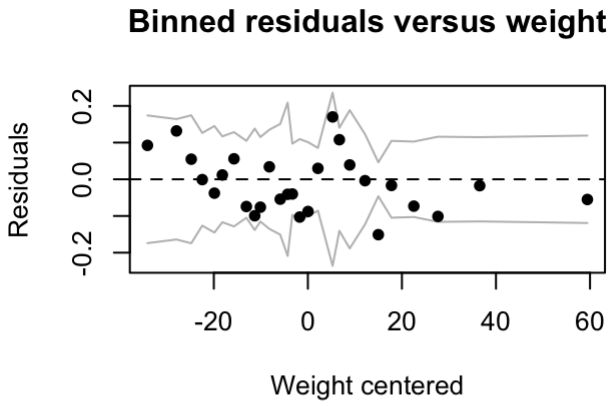
```
##
## Call:
## glm(formula = Premature ~ mage.c + mht.c + mpregwt.c + parity.c +
##     mrace + med + inc, family = binomial, data = smoking)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7479  -0.6782  -0.5586  -0.3936   2.4956
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.539213   1.019718  -0.529  0.59695
## mage.c       0.011029   0.020520   0.537  0.59092
## mht.c       -0.030905   0.042907  -0.720  0.47135
## mpregwt.c   -0.011649   0.005554  -2.098  0.03595 *
## parity.c    -0.004288   0.060336  -0.071  0.94334
## mrace6       0.031751   0.521430   0.061  0.95145
## mrace7       0.774479   0.235676   3.286  0.00102 **
## mrace8       0.749186   0.416061   1.801  0.07176 .
## mrace9      -0.800809   1.056743  -0.758  0.44857
## med1        -0.264090   0.975770  -0.271  0.78666
## med2        -0.704365   0.962162  -0.732  0.46413
## med3        -0.555266   1.009748  -0.550  0.58238
## med4        -1.383927   0.978450  -1.414  0.15724
## med5        -0.943738   0.979773  -0.963  0.33544
## med7         2.022831   1.496298   1.352  0.17641
## inc1        -0.481020   0.513724  -0.936  0.34910
## inc2        -0.623827   0.523303  -1.192  0.23322
## inc3        -0.326230   0.527223  -0.619  0.53607
## inc4        -0.276710   0.536548  -0.516  0.60605
## inc5        -0.166915   0.540957  -0.309  0.75766
## inc6        -0.257112   0.586489  -0.438  0.66110
## inc7        -0.325508   0.539732  -0.603  0.54645
## inc8        -1.438385   1.154583  -1.246  0.21284
## inc9        -0.075186   0.751780  -0.100  0.92034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 841.83  on 868  degrees of freedom
## Residual deviance: 792.87  on 845  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 840.87
##
## Number of Fisher Scoring iterations: 5
```

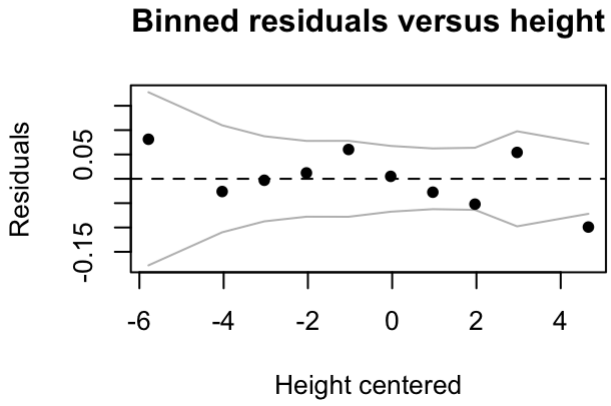
Let's check the binned residuals plots.

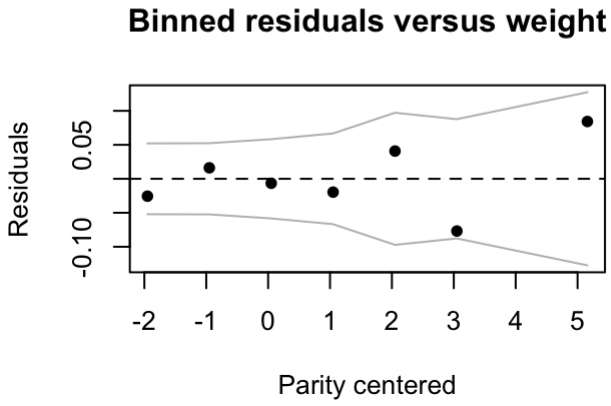
```
par(mfcol=c(2,2))
rawresid1 = smoking$Premature-preterm1$fitted.values
```

```
## Warning in smoking$Premature - preterm1$fitted.values: longer object length
## is not a multiple of shorter object length
```

```
binnedplot(x=smoking$mage.c, y = rawresid1, xlab = "Age centered", ylab = "Residuals", m
ain = "Binned residuals versus age")
binnedplot(x=smoking$mht.c, y = rawresid1, xlab = "Height centered", ylab = "Residuals",
 main = "Binned residuals versus height")
binnedplot(x=smoking$mpregwt.c, y = rawresid1, xlab = "Weight centered", ylab = "Residua
ls", main = "Binned residuals versus weight")
binnedplot(x=smoking$parity.c, y = rawresid1, xlab = "Parity centered", ylab = "Residual
s", main = "Binned residuals versus weight")
```







Residuals fit almost perfectly; no trend observed for weight and height residuals. An attempt to use a quadratic transformation for weight was made to improve the residuals, but the plot has not shown any improvements.

```
tapply(rawresid1, smoking$parity, mean)
```

```
##             0             1             2             3             4
## -0.0255756455  0.0161592751 -0.0065753136 -0.0196317108  0.0408352399
##             5             6             7             8             9
## -0.0768790441 -0.0074703493  0.1414483102 -0.1586000526 -0.0003198194
##            10            11
##  0.7700096353  0.6411778349
```

```
tapply(rawresid1, smoking$mrace, mean)
```

```
##           0           6           7           8           9
## -0.02175704  0.03413195  0.06184951  0.09754114 -0.11731036
```

```
tapply(rawresid1, smoking$med, mean)
```

```
##            0            1            2            3            4
##  0.215703477  0.082120630  0.004789696  0.023452893 -0.066302003
##            5            6            7            8            9
## -0.024059808           NA  0.193137839           NA           NA
```

```
tapply(rawresid1, smoking$inc, mean)
```

```
##             0             1             2             3             4
##  0.0539126222  0.0252228896 -0.0342320075  0.0046940949 -0.0055415458
##             5             6             7             8             9
##  0.0041679603 -0.0011735363  0.0003398662 -0.1236465267 -0.0303706728
```

```
summary(smoking$Premature)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1875  0.0000  1.0000
```

All of the residuals seem to be randomly distributed.

```
#threshold = 0.19
#table(smoking$Premature, preterm1$fitted.values > threshold)
#roc(smoking$Premature, fitted(preterm1), plot=T, legacy.axes=T)

#Due to the technical issues, this part of code was commented. Please see the confusion
 matrix and the specificity curve at the end of the document.
```

For threshold = 0.19, the proportion of “true positive” values is 0.61, which is relatively good, whereas the proportion of “true negative” value is 0.63. These values are approximately equal, so 0.19 is a good choice for the threshold value.

Let’s check some possible interactions with smoke variable since we research the affect of smoke to pre-term births. It would be reasonable to assume some interactions between mother’s age and smoke, smoke and race, smoke and weight and smoke and height.

Let’s explore the interaction between mother’s age and smoke first.

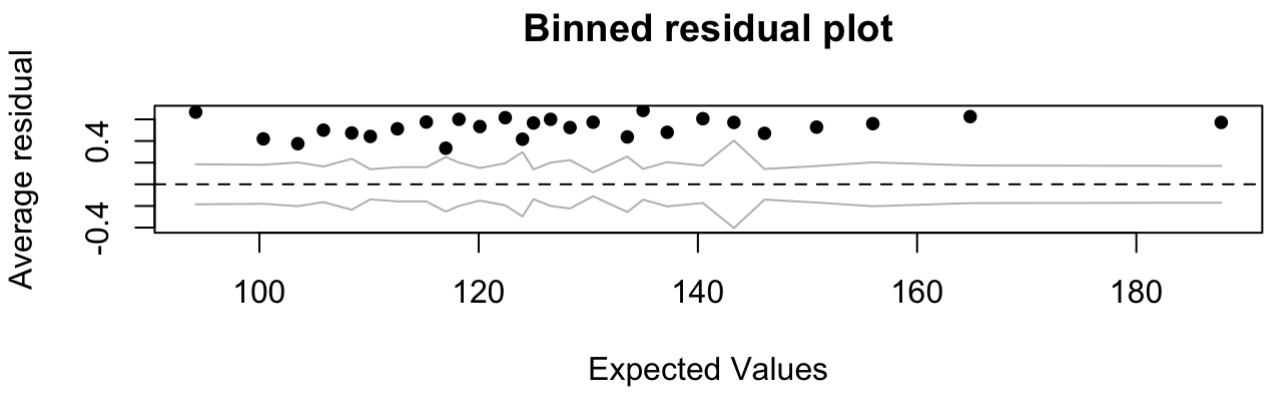
```
par(mfcol=c(2,1))
binnedplot(x = smoking$mage, y = smoking$smoke==0)
binnedplot(x = smoking$mage, y = smoking$smoke==1)
```

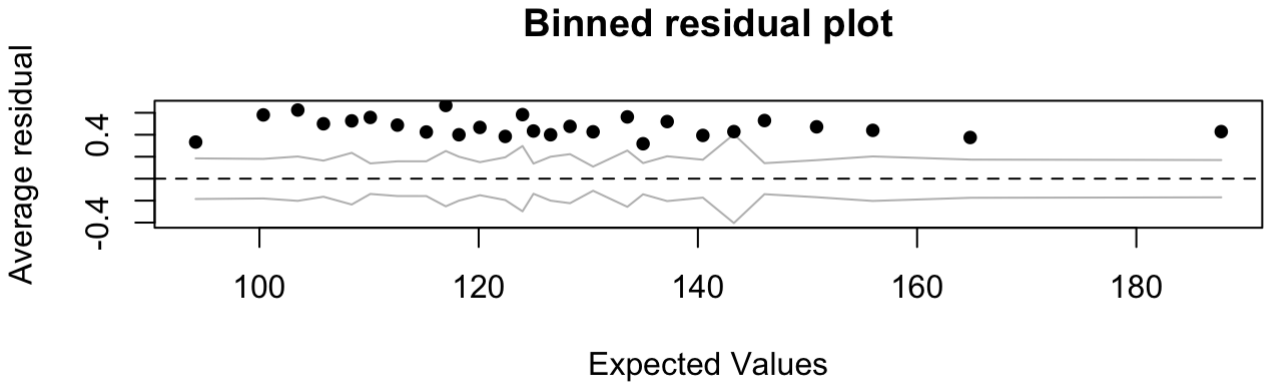

```
preterm2=glm(Premature~ mht.c + mpregwt.c + parity.c + mrace + med + inc + mage.c*smoke,
data = smoking, family = binomial)
anova(preterm1, preterm2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mage.c + mht.c + mpregwt.c + parity.c + mrace + med +
##     inc
## Model 2: Premature ~ mht.c + mpregwt.c + parity.c + mrace + med + inc +
##     mage.c * smoke
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       845     792.87
## 2       843     790.21  2   2.6551   0.2651
```

The interaction between smoke and age is not statistically significant. Let's check the interaction between weight and smoke.

```
par(mfcol=c(2,1))
binnedplot(x = smoking$mpregwt, y = smoking$smoke==0)
binnedplot(x = smoking$mpregwt, y = smoking$smoke==1)
```



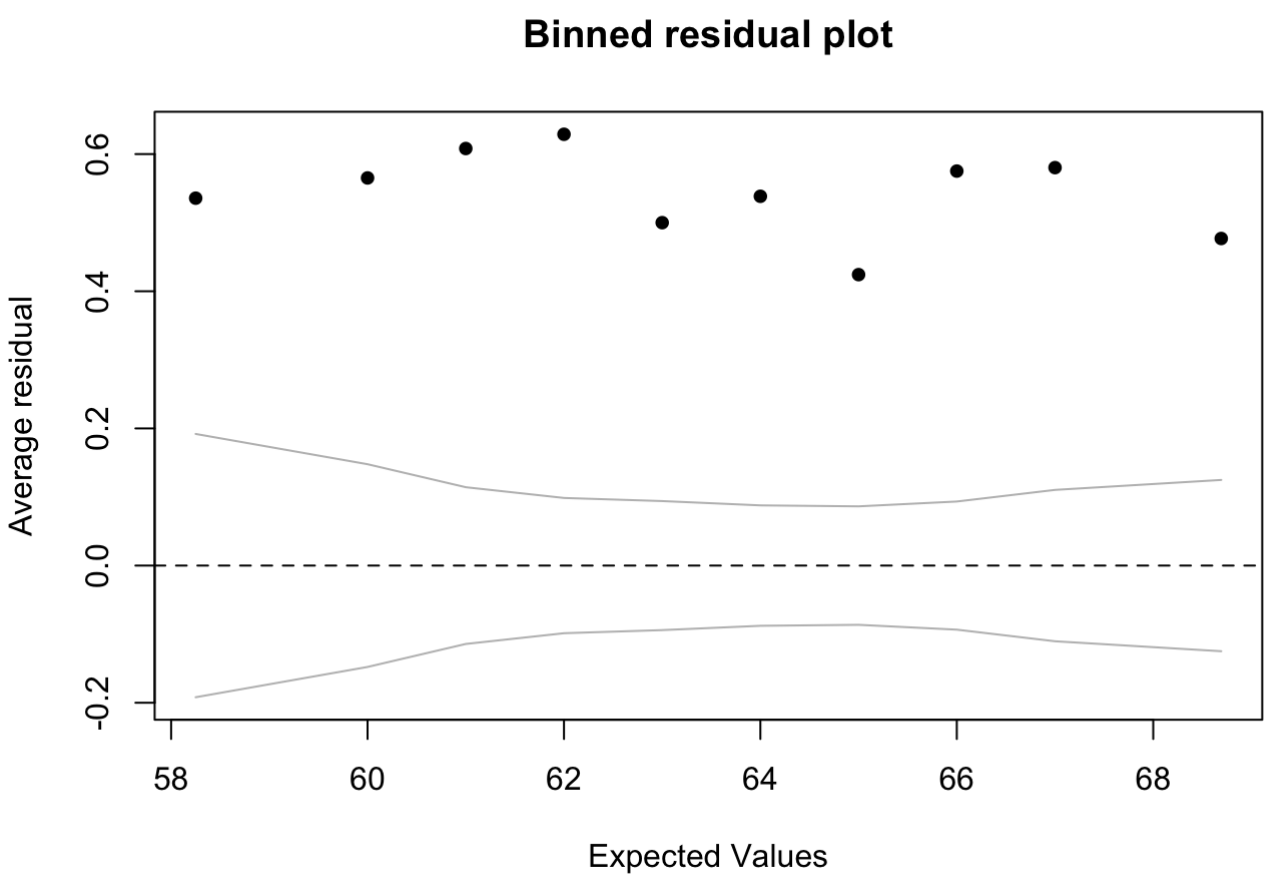



```
preterm3=glm(Premature ~ mage.c + mht.c+ parity.c + mrace + med + inc+mpregwt.c*smoke, d
ata = smoking, family = binomial)
anova(preterm1, preterm3, test="Chisq")
```

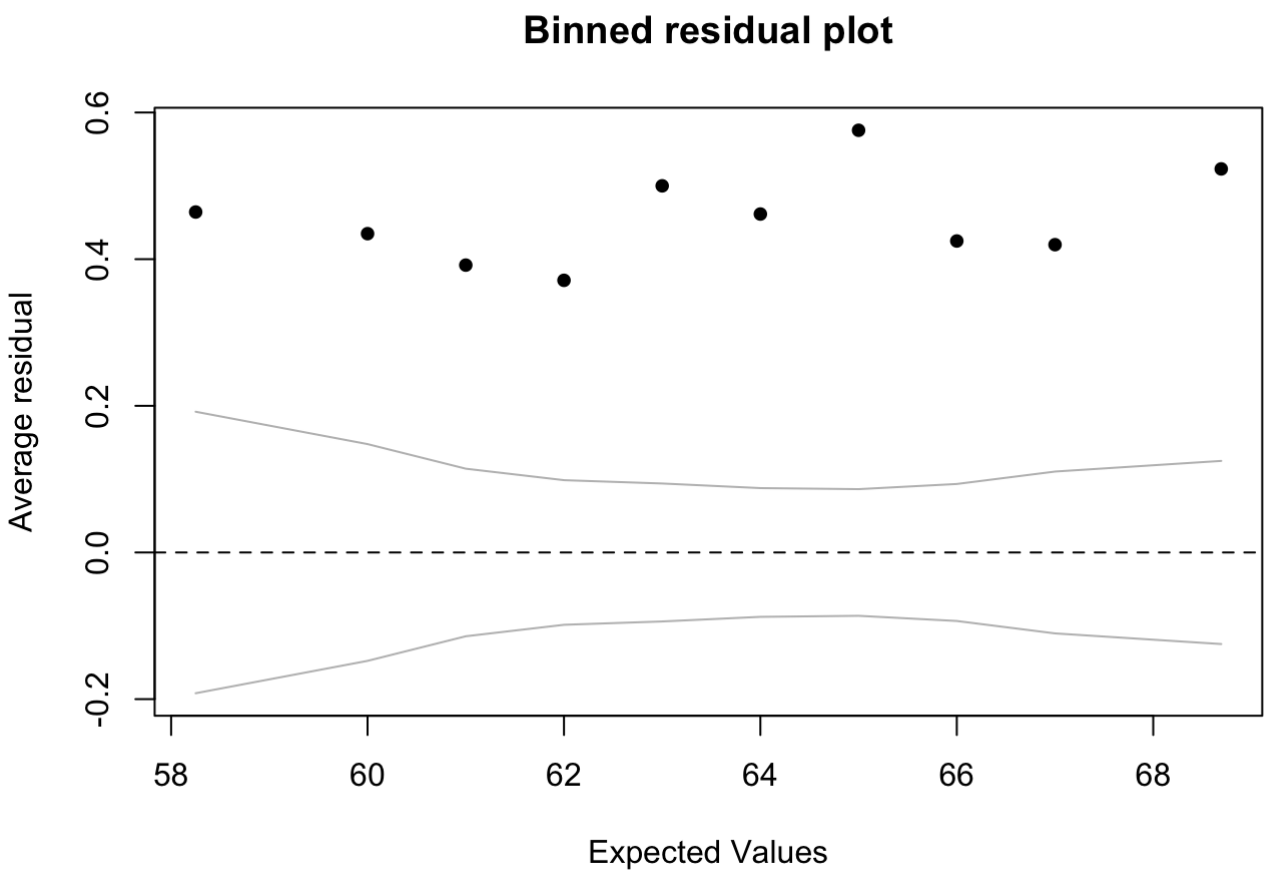
```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mage.c + mht.c + mpregwt.c + parity.c + mrace + med +
##     inc
## Model 2: Premature ~ mage.c + mht.c + parity.c + mrace + med + inc + mpregwt.c *
##     smoke
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       845     792.87
## 2       843     789.28  2   3.5834   0.1667
```

Not a useful interaction as well. Check the interaction between mothers' height and smoking.

```
binnedplot(x = smoking$mht, y = smoking$smoke==0)
```




```
binnedplot(x = smoking$mht, y = smoking$smoke==1)
```




```
preterm4=glm(Premature~mage.c + mpregwt.c + parity.c + mrace + med + inc + mht.c*smoke,
data = smoking, family = binomial)
anova(preterm1, preterm4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mage.c + mht.c + mpregwt.c + parity.c + mrace + med +
##     inc
## Model 2: Premature ~ mage.c + mpregwt.c + parity.c + mrace + med + inc +
##     mht.c * smoke
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       845     792.87
## 2       843     786.65  2   6.2182  0.04464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indeed, this interaction is useful, so we will include it to the model and check if it is improves our model significantly. Lastly, let's check the interaction between mother's race and smoke.

```
par(mfcol=c(2,1))
preterm5=glm(Premature ~ mage.c + mht.c + mpregwt.c + parity.c + med + inc + smoke + mra
ce*smoke, data = smoking, family = binomial)
anova(preterm1, preterm5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mage.c + mht.c + mpregwt.c + parity.c + mrace + med +
##     inc
## Model 2: Premature ~ mage.c + mht.c + mpregwt.c + parity.c + med + inc +
##     smoke + mrace * smoke
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       845     792.87
## 2       840     784.11  5   8.7587   0.1191
```

There are no statistically significant evidences about different effect of smoking on mothers of different races. Also, mother's income and education do not seem to be important predictors for a giving a pre-term birth, so let's do an F-test to check if they should be included to the model.

```
preterm5=glm(Premature ~ mage.c + mpregwt.c + parity.c + mrace + med + mht.c*smoke, data
= smoking, family = binomial)
preterm6=glm(Premature ~ mage.c+ mpregwt.c + parity.c + mrace + inc + mht.c*smoke, data
= smoking, family = binomial)
anova(preterm4, preterm5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mage.c + mpregwt.c + parity.c + mrace + med + inc +
##     mht.c * smoke
## Model 2: Premature ~ mage.c + mpregwt.c + parity.c + mrace + med + mht.c *
##     smoke
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       843     786.65
## 2       852     791.05 -9  -4.4057   0.8827
```

```
anova(preterm4, preterm6, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mage.c + mpregwt.c + parity.c + mrace + med + inc +
##     mht.c * smoke
## Model 2: Premature ~ mage.c + mpregwt.c + parity.c + mrace + inc + mht.c *
##     smoke
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       843     786.65
## 2       849     806.85 -6  -20.198 0.002554 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on results of F-tests, income is not an important variable and we will not include it to the final model, whereas education is important and we will keep it as a predictor.

```
preterm7=glm(Premature ~ mage.c + mpregwt.c +parity.c + mrace + med + mht.c*smoke, data
= smoking, family = binomial)
#threshold = 0.19
#table(smoking$Premature, preterm7$fitted > threshold)
#roc(smoking$Premature, fitted(preterm7), plot=T, legacy.axes=T)

#Due to the technical issues, this part of code was commented. Please see the confusion
 matrix and the specificity curve at the end of the document.
```

Area under the curve now is 0.6646 in comparison with 0.6666 for our first model which did not include interactions. So, including the interaction between the mothers' heights and smoking factor improve the model by 0.2%. It is not a significant impovement, so it depends how essential is to research the effect of interaction. To keep the model simple to interpret, we exclude it from the model.

```
final_model = glm(Premature ~ mage.c + mpregwt.c +parity.c + mrace + med + mht.c + smok
e, data = smoking, family = binomial)
summary(final_model)
```

```
##
## Call:
## glm(formula = Premature ~ mage.c + mpregwt.c + parity.c + mrace +
##     med + mht.c + smoke, family = binomial, data = smoking)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7442  -0.6726  -0.5591  -0.4084   2.4334
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.740126   0.955360  -0.775 0.438512
## mage.c       0.017101   0.019883   0.860 0.389754
## mpregwt.c   -0.011293   0.005521  -2.046 0.040789 *
## parity.c    -0.020333   0.059847  -0.340 0.734040
## mrace6       0.122620   0.523870   0.234 0.814933
## mrace7       0.764865   0.230648   3.316 0.000913 ***
## mrace8       0.827732   0.414961   1.995 0.046073 *
## mrace9      -0.763278   1.054558  -0.724 0.469195
## med1        -0.338929   0.972171  -0.349 0.727367
## med2        -0.723416   0.957668  -0.755 0.450013
## med3        -0.583472   1.003679  -0.581 0.561016
## med4        -1.367568   0.974753  -1.403 0.160620
## med5        -0.925717   0.973168  -0.951 0.341482
## med7         1.946850   1.490985   1.306 0.191638
## mht.c       -0.025777   0.042147  -0.612 0.540793
## smoke0      -0.302499   0.184875  -1.636 0.101790
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 841.83  on 868  degrees of freedom
## Residual deviance: 794.69  on 853  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 826.69
##
## Number of Fisher Scoring iterations: 5
```

```
exp(confint.default(final_model))
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.07334380   3.102926
## mage.c      0.97836797   1.057673
## mpregwt.c   0.97812910   0.999527
## parity.c    0.87142052   1.101821
## mrace6      0.40488752   3.156257
## mrace7      1.36725322   3.376792
## mrace8      1.01452683   5.160540
## mrace9      0.05900268   3.682594
## med1        0.10599657   4.789812
## med2        0.07424304   3.169519
## med3        0.07803122   3.989644
## med4        0.03770174   1.721013
## med5        0.05883066   2.668876
## med7        0.37703131 130.207125
## mht.c       0.89728384   1.058474
## smoke0      0.51435191   1.061676
```

QUESTIONS:

1. Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke? What is a likely range for the difference in odds of pre-term birth for smokers and non-smokers?

Mothers who smoke do have higher chances of pre-term birth. Holding all other variables constant, non-smoking mothers reduce the odds ratio of giving a pre-term birth by the factor of 0.76 with 95% confidence interval (0.53, 1).

2. Is there any evidence that the association between smoking and pre-term birth differs by mother's race? If so, characterize those differences.

We have three times more data about white mothers than data about mothers of any other race. Based on the data we have, we conclude that the interaction between mother's race and smoke is not statistically significant. However, we the probabilities of giving a pre-term birth is slightly different for mothers of different races.

1. Mexican (6): 1.11 95CI (0.39, 3.15)
2. Black (7): 2.11 95CI (1.34, 3.34)
3. Asian (8): 2.52 95CI (1.1, 5.75)
4. Mix (9): 0.49 95CI (0.06, 3.9)

There are not enough evidences to conclude that asian mothers tend to have higher chances of giving a pre-term birth. To make a final conclusion, we have to explore more data with balanced representation of mothers of different races.

3. Are there other interesting associations with the odds of pre-term birth that are worth mentioning?

The interaction between mothers' height and smoking is statistically significant and this fact is confirmed scientifically. Including this interaction to the final model would not improve the results significantly. However, it can be observed that taller smoking mothers tend to have slightly less pre-term birth rates.