Homework 4 Pre-term Birth

Echo Liu

October 10, 2018

Our goal is to study whether maternal smoking would cause higher chances of pre-term birth given outcome variable Premature and predictors like mother's weight and height, mother's race, mother's smoking habit and even father's information. I decided to use cleaned data set which excludes all missing values and all of the variables on the fathers, since some of fathers' variables are missing and can potentially cause multicolinearity problem. Let's first read in the data and look at summary of this data set.

I decided to drop unrelated columns: id and baby weight from the original table. Then I created a new variable called "Premature" to indicate whether mother's gestational age is less than 270 days or nor. Also I renamed race category from number to actual race and grouped mrace from 0-5 into "white".

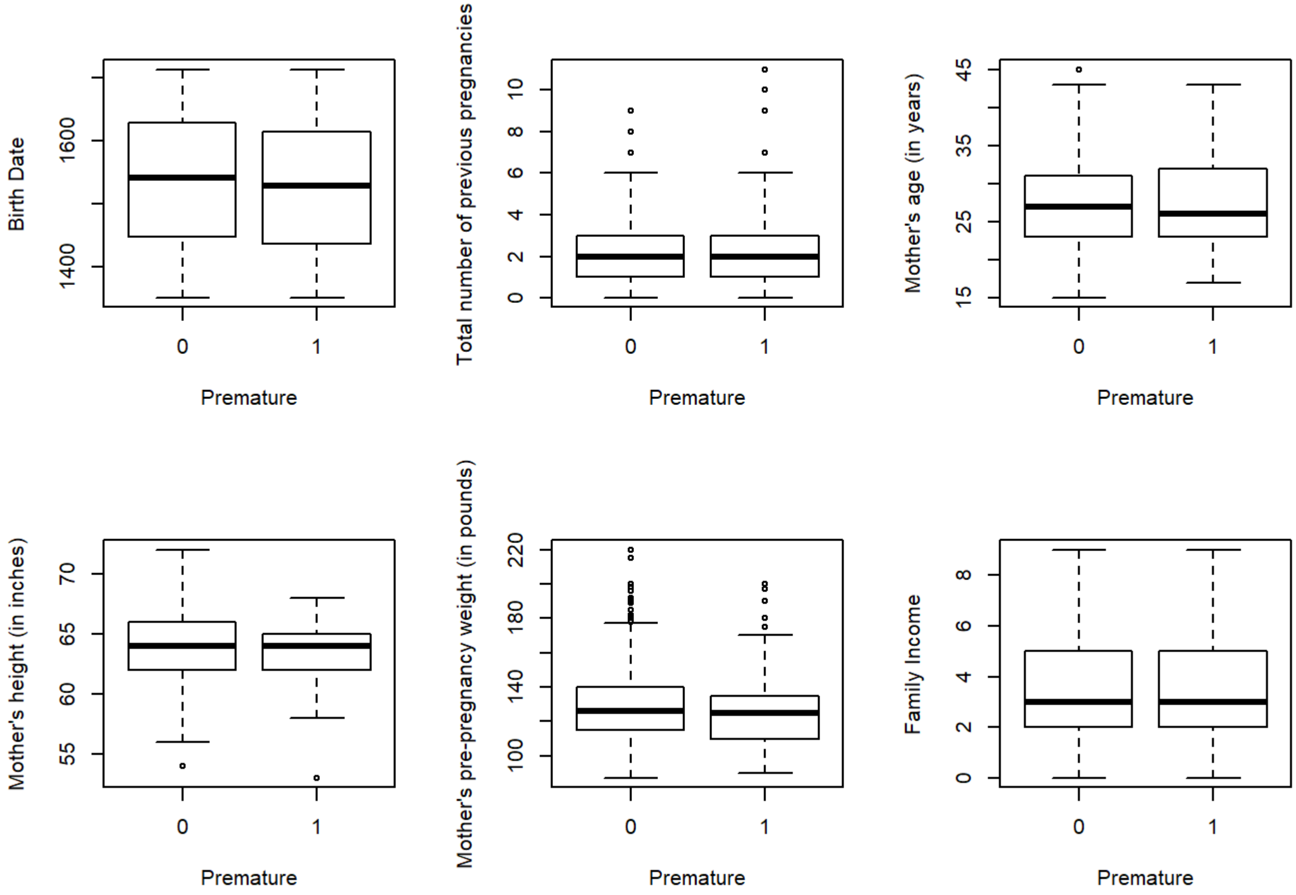
```
#drop unnecessary columns
babies$id = NULL
babies$bwt.oz = NULL

#an indicator variable for Premature (gestational age < 270 days)
n = nrow(babies)
babies$Premature= rep(0,n)
babies$Premature[babies$gestation < 270] = 1

#race data
babies$mracef[0 <= babies$mrace & babies$mrace <= 5] <- "white"
babies$mracef[babies$mrace == 6] <- "mexican"
babies$mracef[babies$mrace == 7] <- "black"
babies$mracef[babies$mrace == 8] <- "asian"
babies$mracef[babies$mrace == 9] <- "mix"
babies$mracef[babies$mrace == 99] <- "unknown"

#smoke data
babies$smokef[babies$smoke == 1] <- "smokes"
babies$smokef[babies$smoke == 0] <- "non-smoke"
```

```
##       date        gestation         parity           mrace      
##  Min.   :1350   Min.   :148.0   Min.   : 0.000   Min.   :0.000  
##  1st Qu.:1444   1st Qu.:272.0   1st Qu.: 1.000   1st Qu.:0.000  
##  Median :1540   Median :279.0   Median : 2.000   Median :2.000  
##  Mean   :1536   Mean   :278.5   Mean   : 1.953   Mean   :2.995  
##  3rd Qu.:1627   3rd Qu.:286.0   3rd Qu.: 3.000   3rd Qu.:7.000  
##  Max.   :1714   Max.   :338.0   Max.   :11.000   Max.   :9.000  
##       mage            med             mht           mpregwt     
##  Min.   :15.00   Min.   :0.000   Min.   :53.00   Min.   : 87.0  
##  1st Qu.:23.00   1st Qu.:2.000   1st Qu.:62.00   1st Qu.:113.0  
##  Median :26.00   Median :2.000   Median :64.00   Median :125.0  
##  Mean   :27.29   Mean   :2.932   Mean   :64.07   Mean   :128.5  
##  3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:66.00   3rd Qu.:140.0  
##  Max.   :45.00   Max.   :7.000   Max.   :72.00   Max.   :220.0  
##       inc            smoke          Premature         mracef         
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Length:869        
##  1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.0000   Class :character  
##  Median :3.000   Median :0.0000   Median :0.0000   Mode  :character  
##  Mean   :3.681   Mean   :0.4638   Mean   :0.1887                     
##  3rd Qu.:5.000   3rd Qu.:1.0000   3rd Qu.:0.0000                     
##  Max.   :9.000   Max.   :1.0000   Max.   :1.0000                     
##     smokef         
##  Length:869        
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

Mean of Premature is 0.19, which means premature birth cases only accounts for 19% of all data points, making us hard to do accurate predictions. ### Exploratory Data Analysis

Box Plots for Continuous Variables

Since family income is categorized almost linearly (except for category 9 which includes family earning more than 15000) and there is no unknown or unasked data point in this case, we could treat “inc” as a continous variable.



Tabular Format for Categorical Variables

```
tapply(babies$Premature, babies$med, mean)
```

```
##         0         1         2         3         4         5         7 
## 0.4000000 0.2769231 0.1900312 0.2340426 0.1182266 0.1698113 0.7500000
```

```
table(babies$med)
```

```
## 
##   0   1   2   3   4   5   7 
##   5 130 321  47 203 159   4
```

```
#The large uncertainties occurs at education= 0 or 7 which is due to lack of observations in those two cases.

tapply(babies$Premature,babies$mracef, mean)
```

```
##      asian      black    mexican        mix      white 
## 0.32352941 0.26627219 0.24000000 0.06666667 0.16134185
```

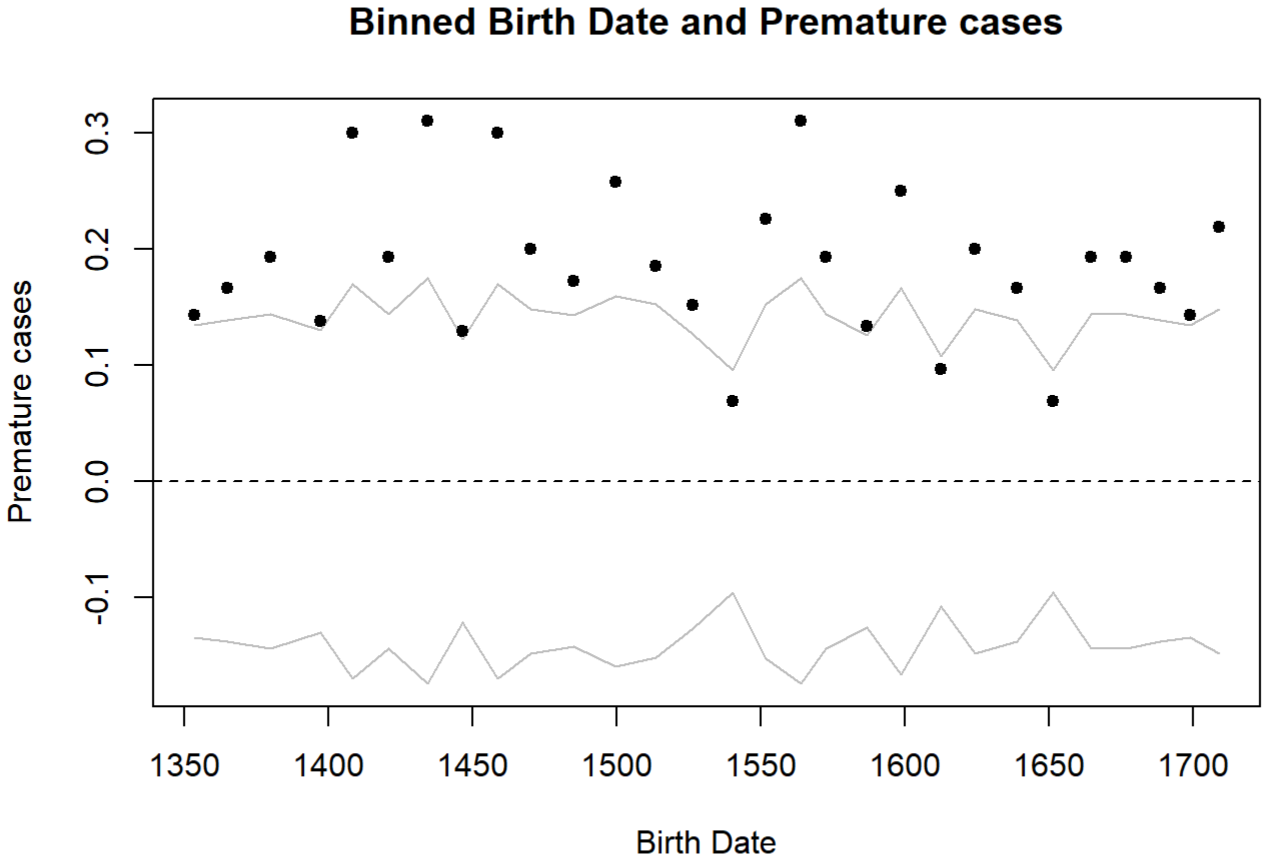
```
table(babies$mracef)
```

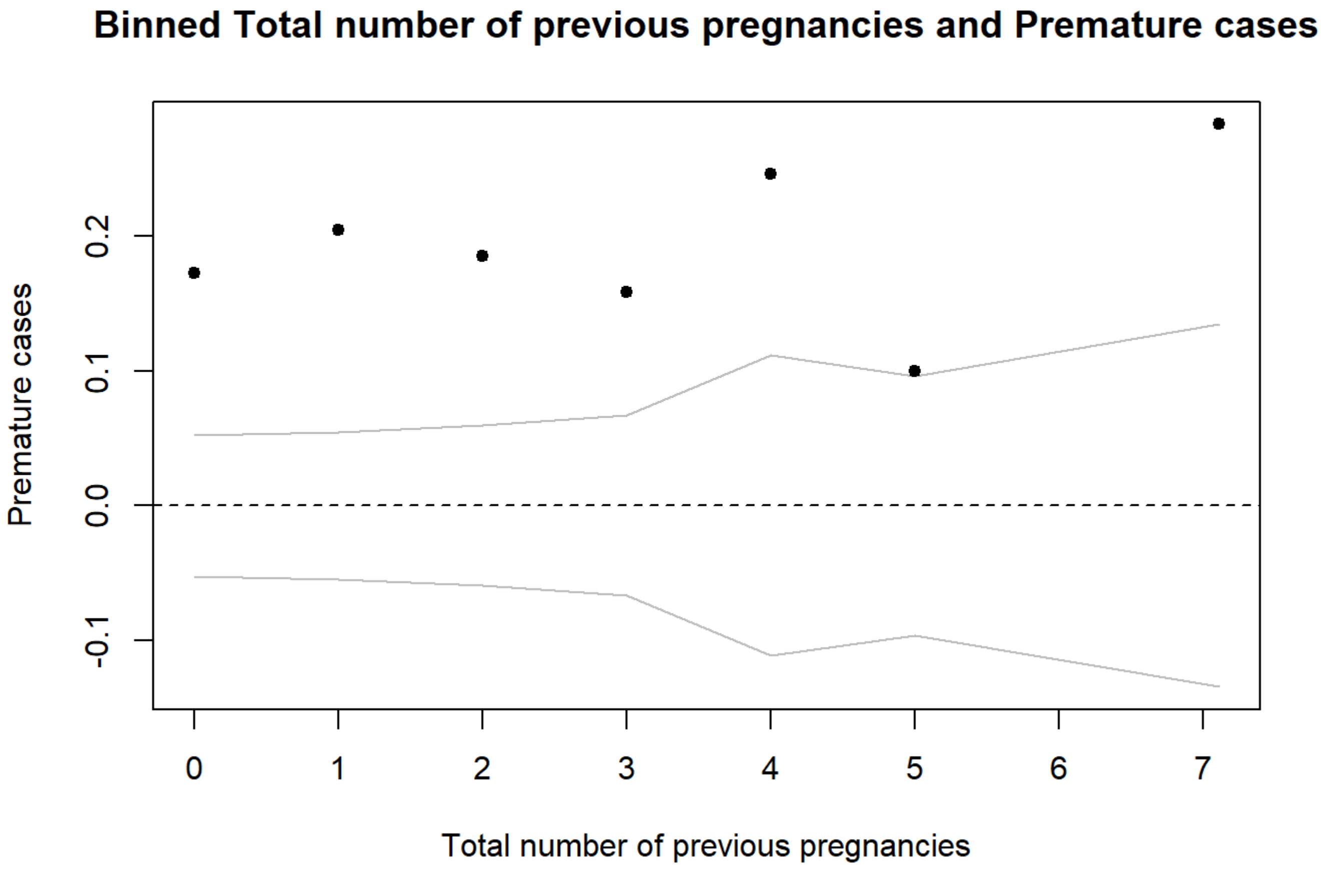
```
##
##   asian   black mexican     mix   white
##      34     169      25      15     626
```

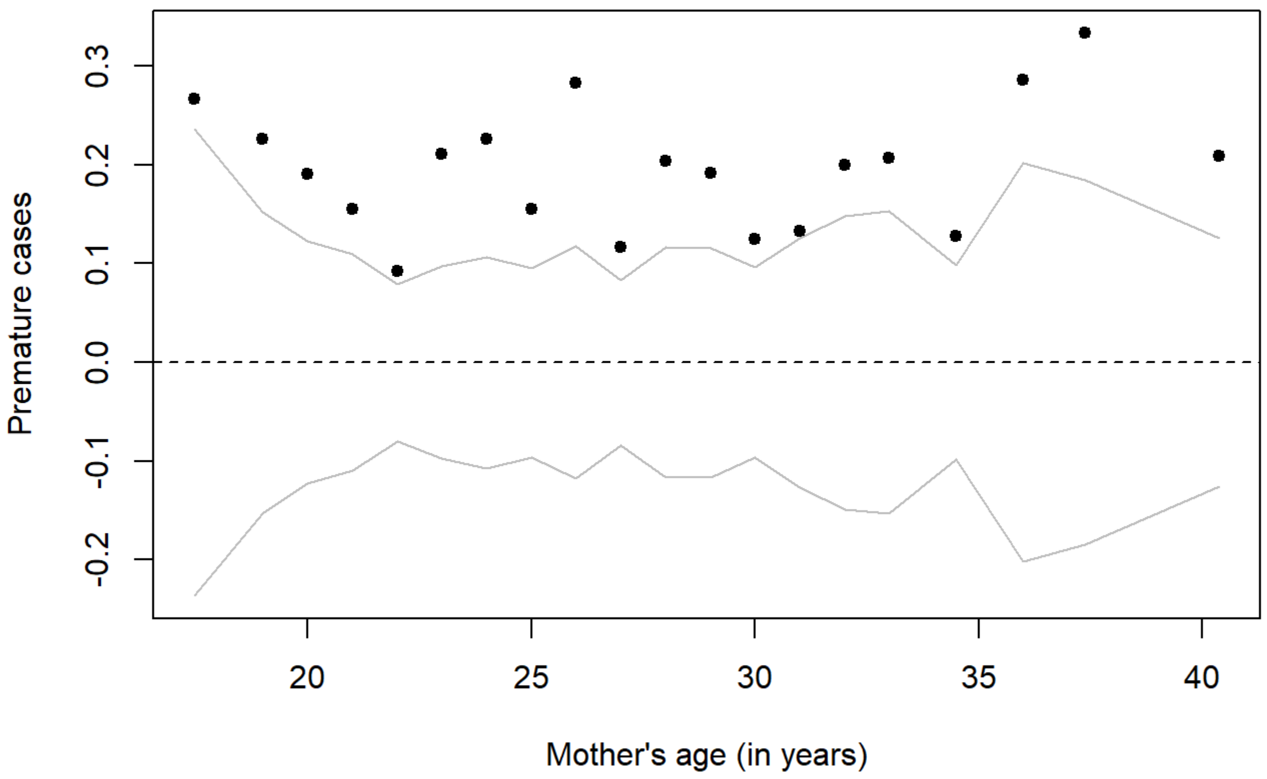
```
#When mother's race is "mix", there are few observations which also explains why mean of Prematu
re for mix is so low. Also, notice that majority of data points are collected from white mother,
 therefore, it's reasonable to make "white" as a baseline instead of other races.

tapply(babies$Premature, babies$smokef, mean)
```

```
## non-smoke    smokes
## 0.1652361 0.2158809
```

Binned plots of Continuous Variables versus Premature

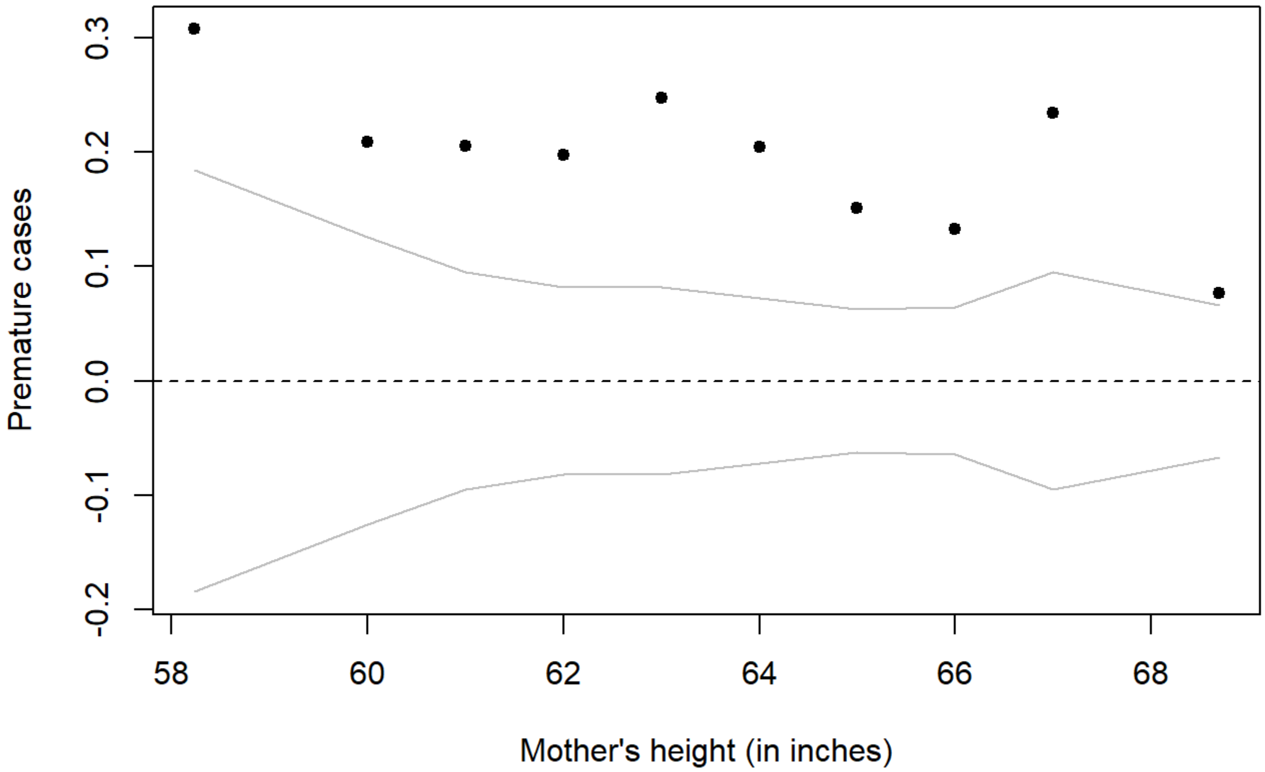

Binned Birth Date and Premature cases
Premature cases
0.3
0.2
0.1
0.0
-0.1
1350
1400
1450
1500
1550
1600
1650
1700
Birth Date


Binned Total number of previous pregnancies and Premature cases
Premature cases
0.2
0.1
0.0
-0.1
0
1
2
3
4
5
6
7
Total number of previous pregnancies

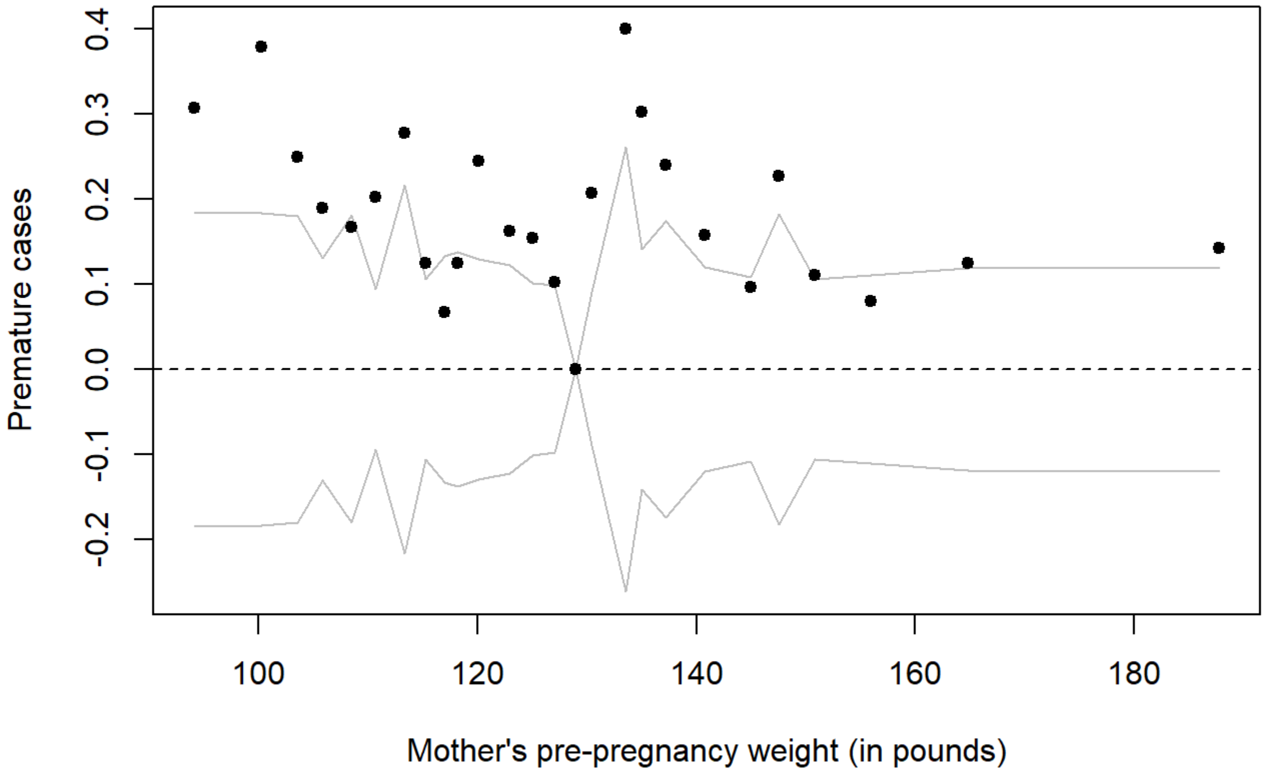


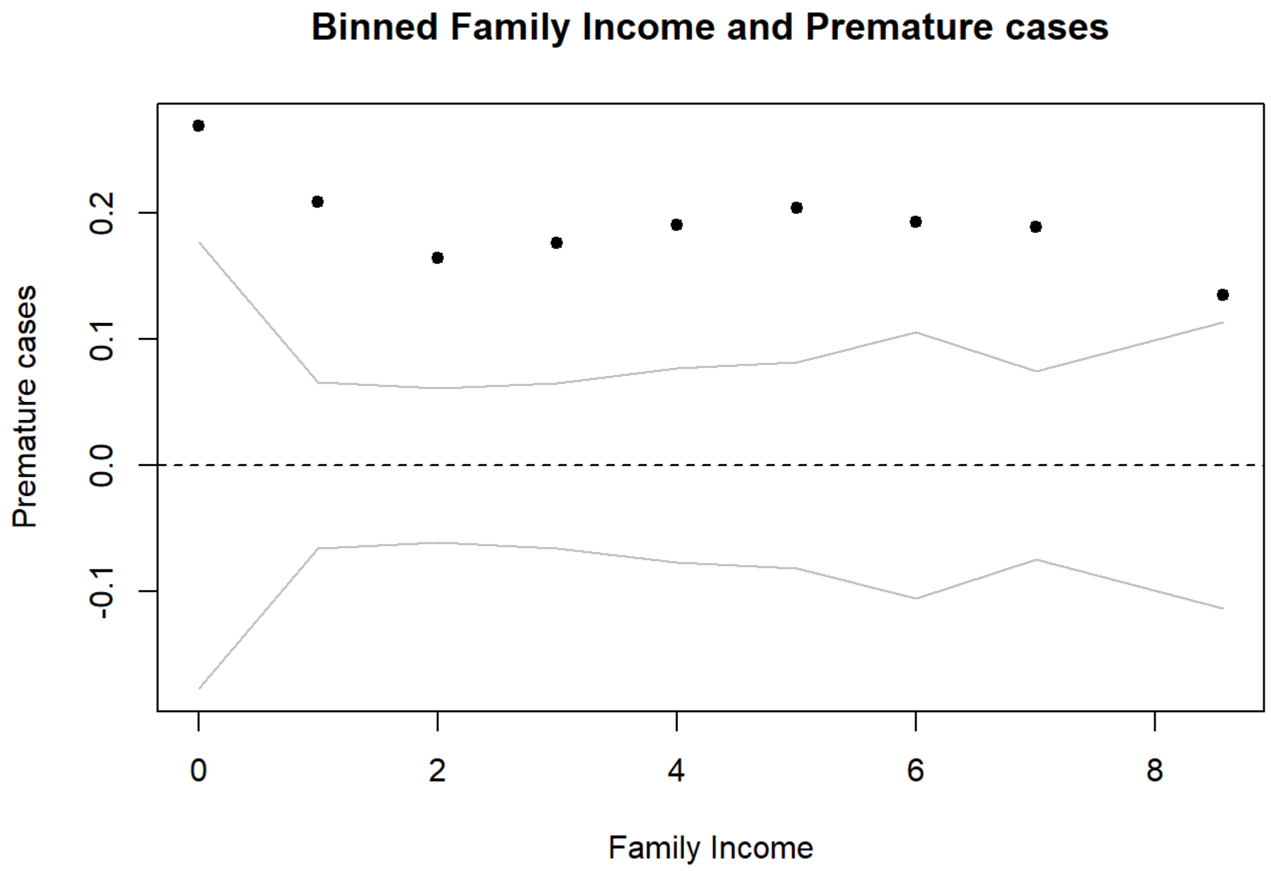
Binned Mother's age (in years) and Premature cases

Binned Mother's height (in inches) and Premature cases



Binned Mother's pre-pregnancy weight (in pounds) and Premature case





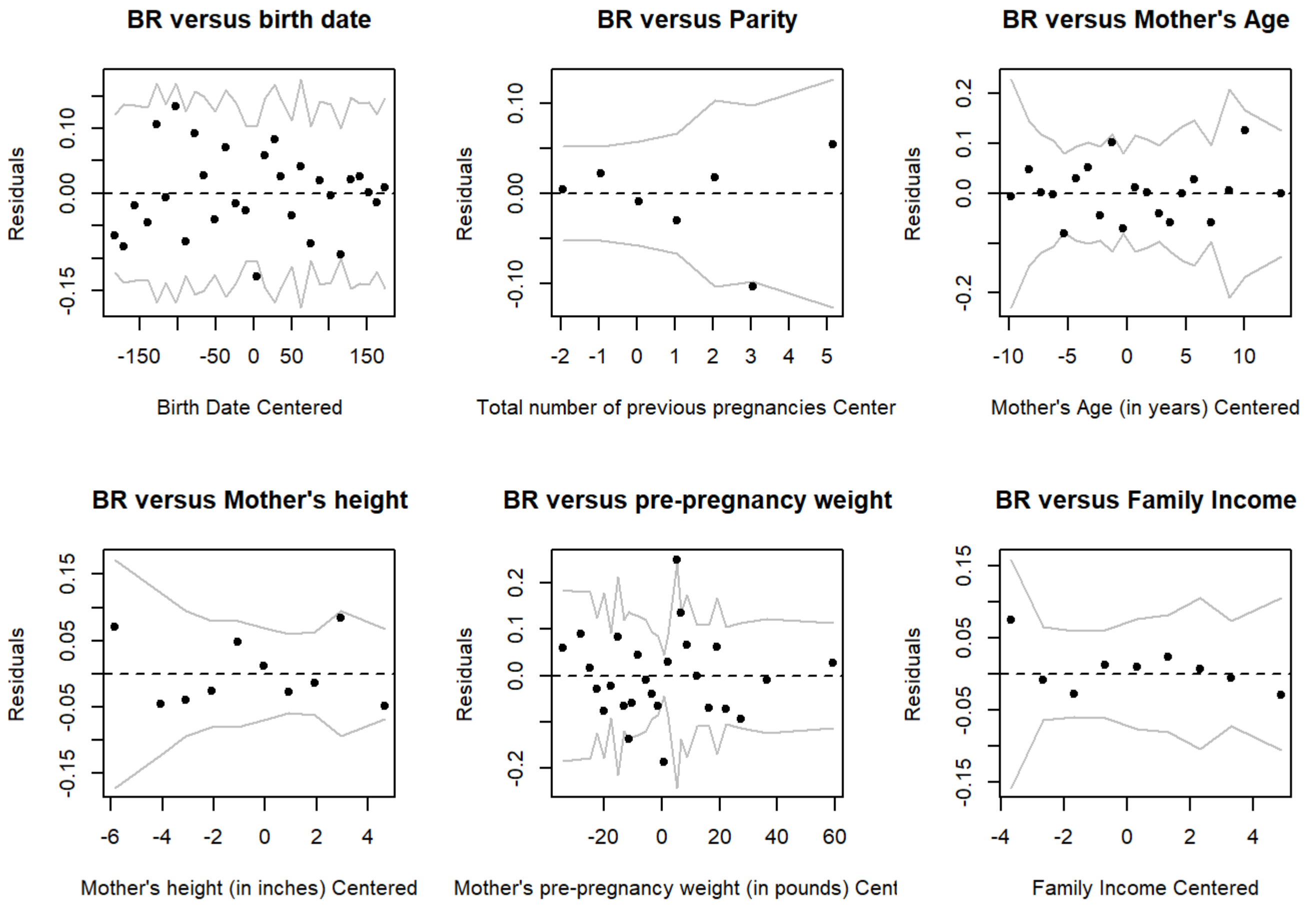
No real patterns show up in the scatter plots, so there is no obvious transformation suggested.

Model1: Plain vanilla with mean centering

```
#Let's try a logistic regression that has a main effect for every variable and linear predictor
s. Begin by centering the continuous predictor.
babies$date.c = babies$date - mean(babies$date)
babies$parity.c = babies$parity - mean(babies$parity)
babies$mage.c = babies$mage - mean(babies$mage)
babies$mht.c = babies$mht - mean(babies$mht)
babies$mpregwt.c = babies$mpregwt - mean(babies$mpregwt)
babies$inc.c = babies$inc - mean(babies$inc)

babiesreg1 = glm(Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c + inc.c + as.factor
(med) + relevel(as.factor(mracef),ref = "white") +as.factor(smoke), data = babies, family = bino
mial)
```

Model Diagnostic

Part A: Binned Residuals



Not as much of a trend for those plots.

Average Residuals by categorical variables

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

```
##             0             1             2             3             4 
## -4.807266e-15 -2.778119e-13 -1.936260e-13 -2.329691e-13 -3.026863e-13 
##             5             7 
## -1.015243e-16 -4.163336e-16
```

```
tapply(rawresid1, babies$mracef, mean)
```

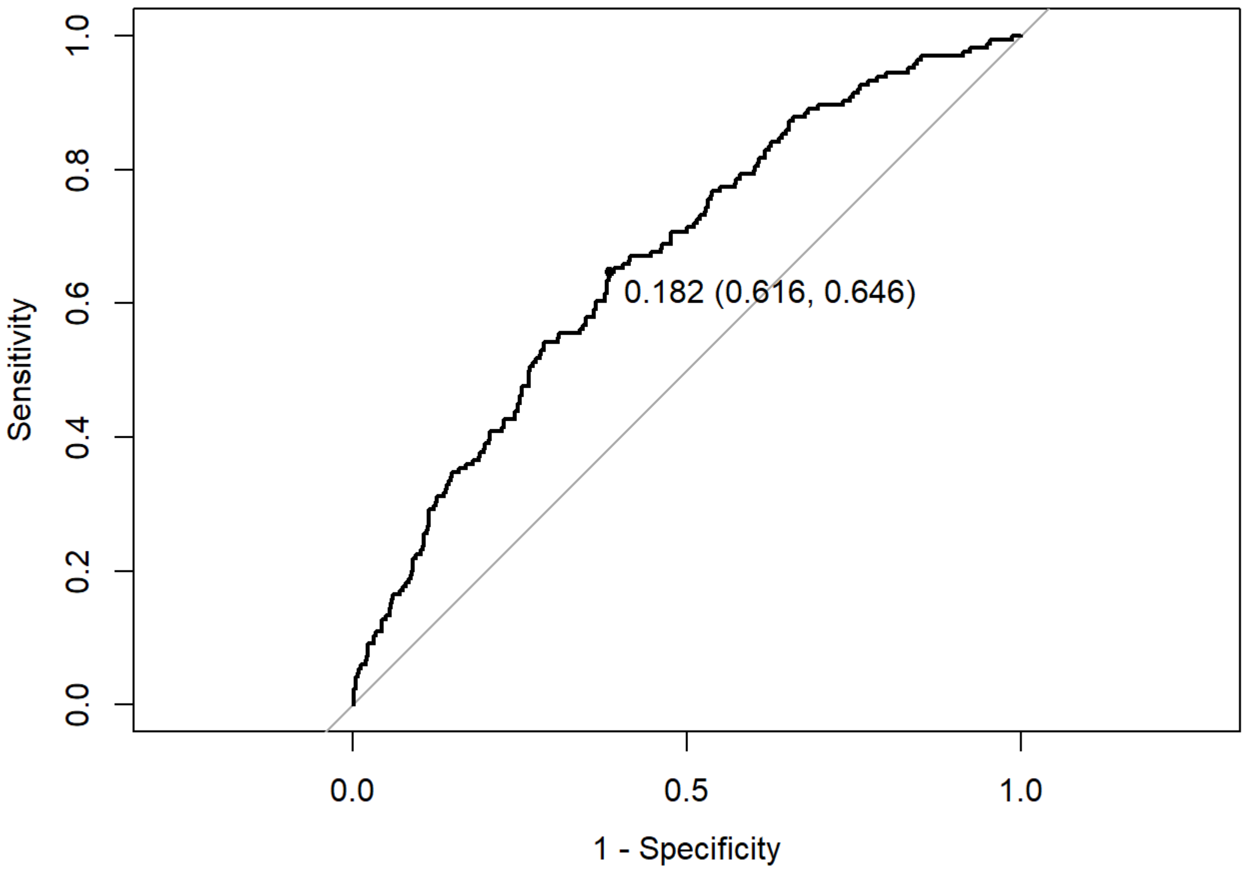
```
##         asian         black       mexican           mix         white 
## -6.283749e-17 -4.187795e-16 -3.458150e-16 -1.136372e-11 -2.698939e-16
```

```
tapply(rawresid1, babies$smokef, mean)
```

```
##     non-smoke        smokes 
## -2.805373e-13 -9.919669e-14
```

Part B: ROC Curve and Confusion Matrix

Let’s sketch ROC plot first to find the optimum threshold.



```
##
## Call:
## roc.default(response = babies$Premature, predictor = fitted(babiesreg1),     plot = T, print.
thres = "best", legacy.axes = T)
##
## Data: fitted(babiesreg1) in 705 controls (babies$Premature 0) < 164 cases (babies$Premature
1).
## Area under the curve: 0.6667
```

The ROC curve is pretty tight to the line, thus it’s not a strongly predictive logistic regression. Area under the curve is 0.6667. The true positive rate is 0.646 and false negative rate is 0.616, which means false positive rate is 0.384. We have relatively high value of sensitivity and relatively low value of specificity. It’s already a model with decent accuracy given a small premature data set.

Next, let’s do the confusion matrix with 0.182 threshold.

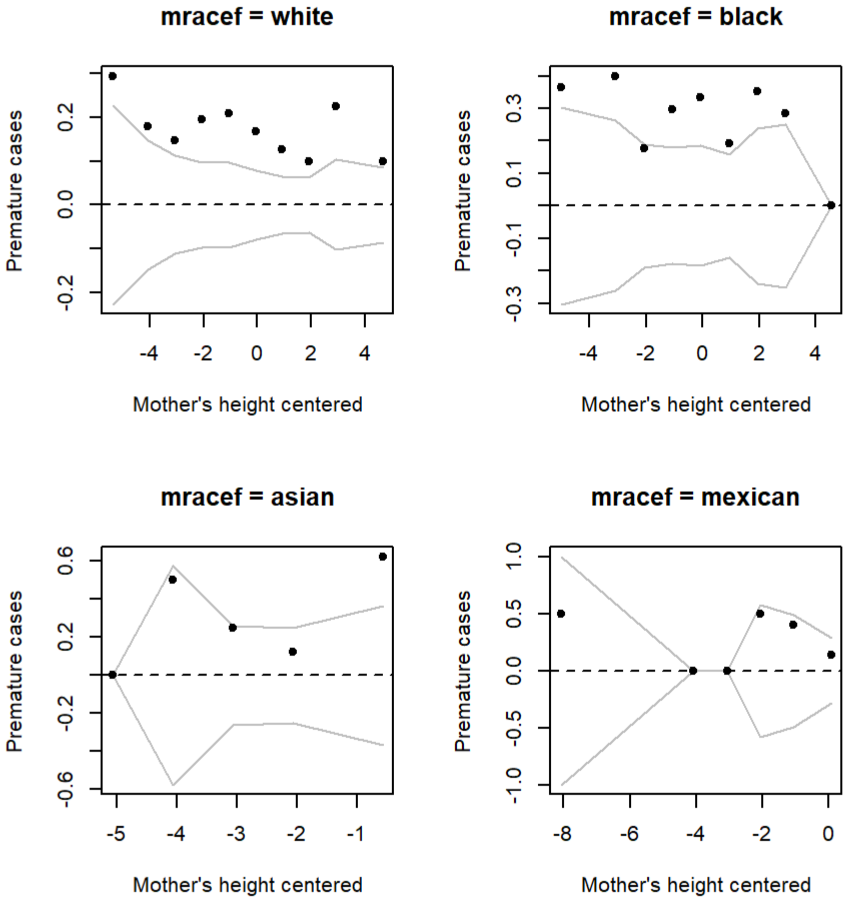
```
threshold = 0.182
table(babies$Premature, babiesreg1$fitted.values > threshold)
```

```
##
##     FALSE TRUE
##   0   433  272
##   1    58  106
```

The confusion matrix again shows that mis-classification rate is relatively low.

Explore Interactions

I'll explore some interactions which would provide scientifically meaningful explanations. We can get an idea about whether those interactions exist by plotting binned plot. Let's first see if there are any interactions between mother's height and mother's race.



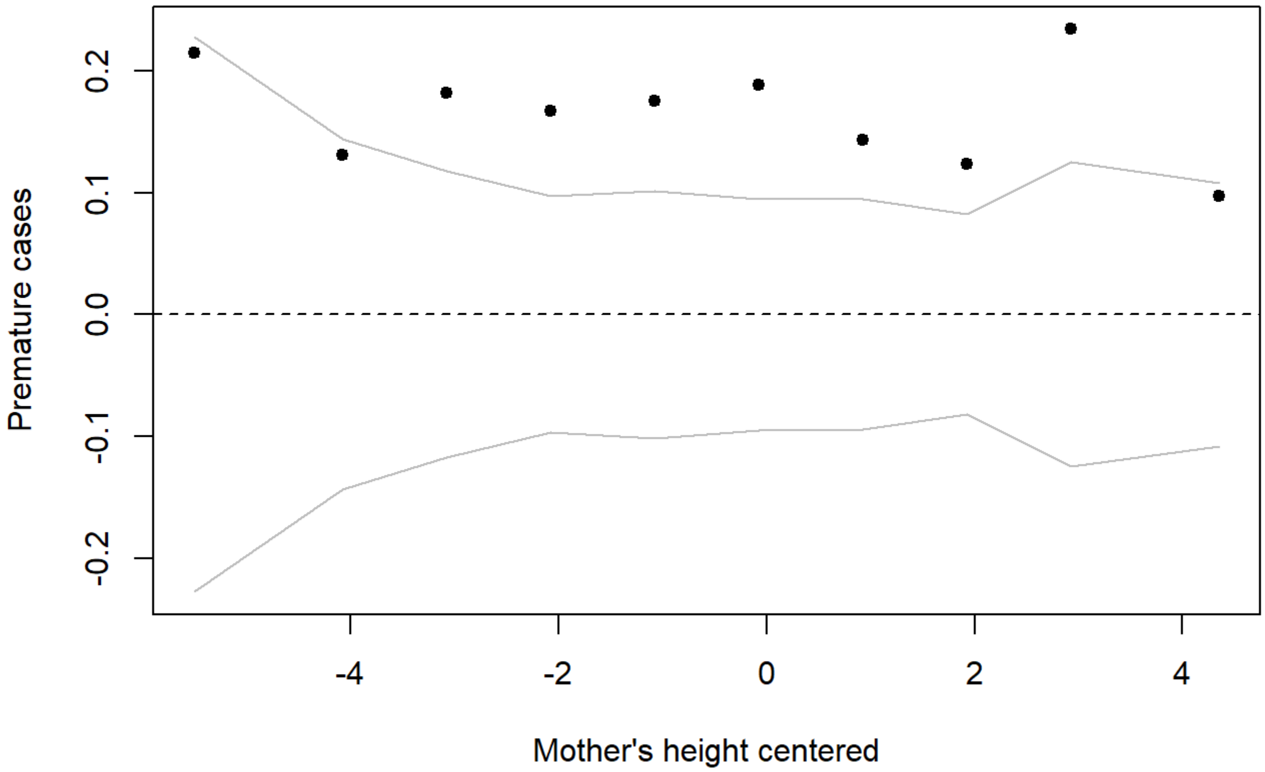
Since binned plot corresponding to each race has different shapes, we assume that we don't have enough data in each category to observe any pattern. But before quitting, let's do a change in deviance test to decide whether this interaction is significant.

```
babiesreg_1 = glm(Premature ~ date.c + parity.c + mage.c + mht.c * relevel(as.factor(mracef),ref
= "white") +  mpregwt.c + inc.c + as.factor(med) + as.factor(smokef), data = babies, family = bi
nomial)
anova(babiesreg_1, babiesreg1, test = "Chisq")
```

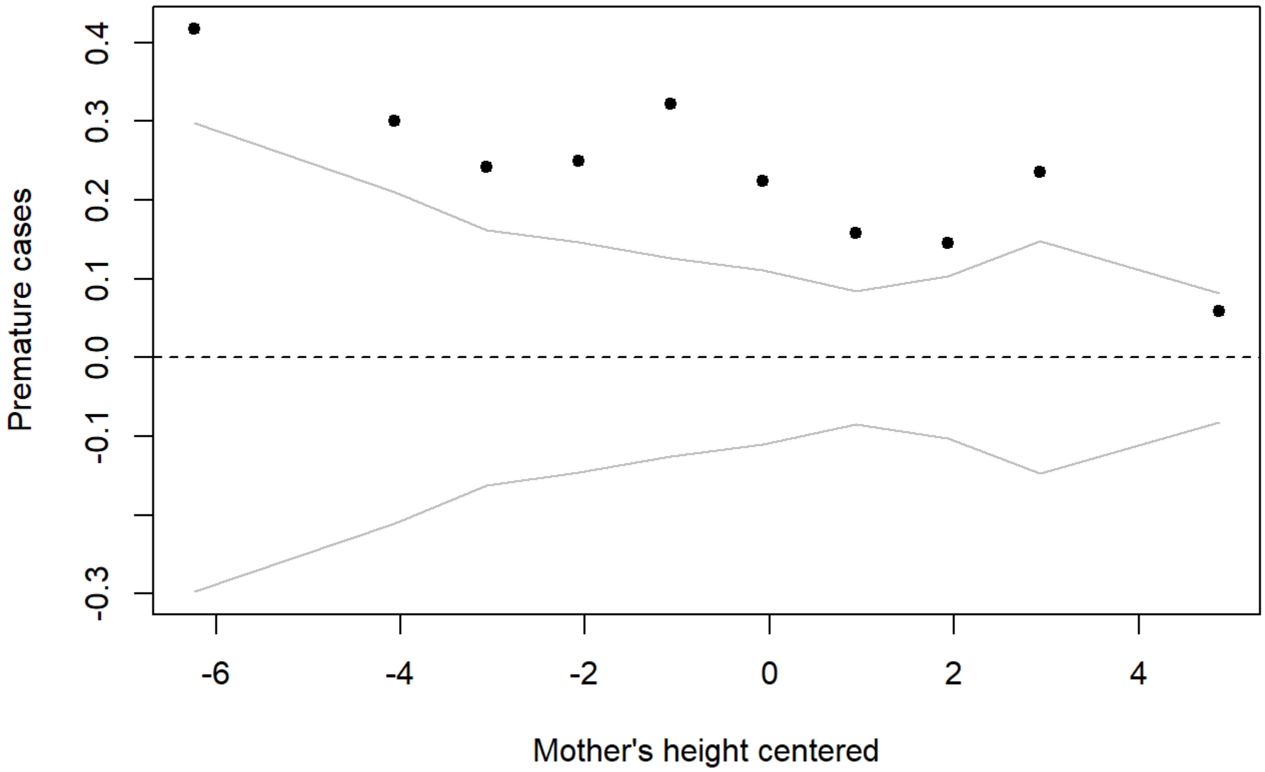
```
## Analysis of Deviance Table
##
## Model 1: Premature ~ date.c + parity.c + mage.c + mht.c * relevel(as.factor(mracef),
##     ref = "white") + mpregwt.c + inc.c + as.factor(med) + as.factor(smokef)
## Model 2: Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c + as.factor(med) + relevel(as.factor(mracef), ref = "white") +
##     as.factor(smoke)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       847     789.54
## 2       851     793.35 -4  -3.8149   0.4316
```

Since p-value of this interaction is 0.4316,mht.c * mracef seems to be a useless predictor. Secondly, let’s then try the interaction between smoke and heights.

Binned Mother's height and Premature cases (smokef == nonsmoke)



Binned Mother's height and Premature cases (smokef == smokes)



It seems that when mother is a non-smoker, binned plot of Premature vs Mother's height seems relatively flat whereas when mother is a smoker, there is a tendency that as mother's height increases, the chance of pre-term birth decreases. Let's do a change in deviance test to check whether this is a useful predictor.

```
babiesreg2 = glm(Premature ~ date.c + parity.c + mage.c + mht.c * as.factor(smokef) +  mpregwt.c
+ inc.c + as.factor(med) + relevel(as.factor(mracef),ref = "white"), data = babies, family = bin
omial)

anova(babiesreg2, babiesreg1, test = "Chisq")
```

```
## Analysis of Deviance Table
## 
## Model 1: Premature ~ date.c + parity.c + mage.c + mht.c * as.factor(smokef) + 
##     mpregwt.c + inc.c + as.factor(med) + relevel(as.factor(mracef), 
##     ref = "white")
## Model 2: Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c + 
##     inc.c + as.factor(med) + relevel(as.factor(mracef), ref = "white") + 
##     as.factor(smoke)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1       850     789.79                       
## 2       851     793.35 -1  -3.5631  0.05908 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Change of deviance test shows that p-value of this interaction is 0.05908, which is not bad. However, I don't know if there is any scientific interpretation for relationship between chance of pre-term birth and mother's height given whether mother smokes or not. It seems to me that including this interaction is merely overfitting the model. I decide to keep the baseline model at the end.

Interpretation

```
## 
## Call:
## glm(formula = Premature ~ date.c + parity.c + mage.c + mht.c + 
##     mpregwt.c + inc.c + as.factor(med) + relevel(as.factor(mracef), 
##     ref = "white") + as.factor(smoke), family = binomial, data = babies)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7079  -0.6710  -0.5541  -0.4058   2.4656  
## 
## Coefficients:
##                                                     Estimate Std. Error
## (Intercept)                                      -1.0416968  0.9624726
## date.c                                           -0.0009139  0.0008539
## parity.c                                         -0.0180399  0.0598359
## mage.c                                            0.0155956  0.0205714
## mht.c                                            -0.0300505  0.0424108
## mpregwt.c                                        -0.0111238  0.0055214
## inc.c                                             0.0223035  0.0431375
## as.factor(med)1                                  -0.3014318  0.9756556
## as.factor(med)2                                  -0.7232492  0.9623161
## as.factor(med)3                                  -0.6184057  1.0093460
## as.factor(med)4                                  -1.3795497  0.9789037
## as.factor(med)5                                  -0.9485027  0.9802987
## as.factor(med)7                                   1.9384074  1.4904691
## relevel(as.factor(mracef), ref = "white")asian    0.8076455  0.4161622
## relevel(as.factor(mracef), ref = "white")black    0.7857279  0.2327918
## relevel(as.factor(mracef), ref = "white")mexican  0.1492711  0.5238502
## relevel(as.factor(mracef), ref = "white")mix     -0.7557067  1.0566729
## as.factor(smoke)1                                 0.2818801  0.1857989
##                                                  z value Pr(>|z|)    
## (Intercept)                                       -1.082 0.279113    
## date.c                                            -1.070 0.284503    
## parity.c                                          -0.301 0.763041    
## mage.c                                             0.758 0.448378    
## mht.c                                             -0.709 0.478599    
## mpregwt.c                                         -2.015 0.043940 *  
## inc.c                                              0.517 0.605133    
## as.factor(med)1                                   -0.309 0.757357    
## as.factor(med)2                                   -0.752 0.452309    
## as.factor(med)3                                   -0.613 0.540088    
## as.factor(med)4                                   -1.409 0.158752    
## as.factor(med)5                                   -0.968 0.333262    
## as.factor(med)7                                    1.301 0.193418    
## relevel(as.factor(mracef), ref = "white")asian     1.941 0.052295 .  
## relevel(as.factor(mracef), ref = "white")black     3.375 0.000738 ***
## relevel(as.factor(mracef), ref = "white")mexican   0.285 0.775682    
## relevel(as.factor(mracef), ref = "white")mix      -0.715 0.474501    
## as.factor(smoke)1                                  1.517 0.129235    
## ---
## 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: 793.35  on 851  degrees of freedom
## AIC: 829.35
##
## Number of Fisher Scoring iterations: 5
```

```
##                                                          2.5 %     97.5 %
## (Intercept)                                         0.05349814   2.327314
## date.c                                              0.99741587   1.000760
## parity.c                                            0.87343974   1.104327
## mage.c                                              0.97557956   1.057508
## mht.c                                               0.89299521   1.054507
## mpregwt.c                                           0.97829346   0.999698
## inc.c                                               0.93965319   1.112769
## as.factor(med)1                                     0.10929751   5.006906
## as.factor(med)2                                     0.07358205   3.199056
## as.factor(med)3                                     0.07452002   3.895711
## as.factor(med)4                                     0.03695088   1.714406
## as.factor(med)5                                     0.05670727   2.645467
## as.factor(med)7                                     0.37423985 128.982051
## relevel(as.factor(mracef), ref = "white")asian   0.99201468   5.069836
## relevel(as.factor(mracef), ref = "white")black   1.39022329   3.462502
## relevel(as.factor(mracef), ref = "white")mexican 0.41583933   3.241378
## relevel(as.factor(mracef), ref = "white")mix     0.05920510   3.725996
## as.factor(smoke)1                                   0.92101495   1.907969
```

Intercept Odds of pre-term birth for a baby who is born on the 441st day counting from January 1, 1961, given birth by a white non-smoking non-education mother with mean parity, age, heigh, pre-pregnancy weight, who comes from a mean-income family, is $e^{-1.0416968} \approx 0.3528$ (95% CI:0.05349814, 2.327314) .

Mother’s Race (and all other categorical variables could be interpreted the same way) If we keep all other predictors constant and we change mother’s race from White to Black, odds of pre-term birth for baby will increase by $e^{0.7857} \approx 2.2$ (95% CI: 1.3902, 3.4625)

Mother’s Height (and all other continuous variables could be interpreted the same way) If we keep all other predictors constant and increase mother’s height by 1 unit, odds of pre-term birth will increase by $e^{-0.0301} \approx 0.97$ (95% CI: 0.893, 1.055).

Question 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?

We check the significance of smokef variable through a change in deviance test.

```
babiesreg3 = glm(Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c + inc.c + as.factor
(med) + relevel(as.factor(mracef),ref = "white") , data = babies, family = binomial)

anova(babiesreg3, babiesreg1,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c + as.factor(med) + relevel(as.factor(mracef), ref = "white")
## Model 2: Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c + as.factor(med) + relevel(as.factor(mracef), ref = "white") +
##     as.factor(smoke)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       852     795.66
## 2       851     793.35  1   2.3066   0.1288
```

P-value for smokef is 0.1288, so smoke doesn't seem to be a useful predictor. But due to its scientific value in this experiment, we still retain this predictor.

What is a likely range for the difference in odds of pre-term birth for smokers and non-smokers?

```
##                                                          2.5 %     97.5 %
## (Intercept)                                         0.05349814   2.327314
## date.c                                              0.99741587   1.000760
## parity.c                                            0.87343974   1.104327
## mage.c                                              0.97557956   1.057508
## mht.c                                               0.89299521   1.054507
## mpregwt.c                                           0.97829346   0.999698
## inc.c                                               0.93965319   1.112769
## as.factor(med)1                                     0.10929751   5.006906
## as.factor(med)2                                     0.07358205   3.199056
## as.factor(med)3                                     0.07452002   3.895711
## as.factor(med)4                                     0.03695088   1.714406
## as.factor(med)5                                     0.05670727   2.645467
## as.factor(med)7                                     0.37423985 128.982051
## relevel(as.factor(mracef), ref = "white")asian     0.99201468   5.069836
## relevel(as.factor(mracef), ref = "white")black     1.39022329   3.462502
## relevel(as.factor(mracef), ref = "white")mexican 0.41583933   3.241378
## relevel(as.factor(mracef), ref = "white")mix       0.05920510   3.725996
## as.factor(smoke)1                                   0.92101495   1.907969
```

From the confidence interval summary, we're 95% confident that for mother who smokes, the odds of pre-term birth is expected to increase $e^{0.281881} \approx 1.3256$ (95%CI:0.92101495, 1.907969) when compared to mother who doesn't smoke holding all else constant.

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

```
babiesreg4 = glm(Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c + inc.c + as.factor
(med) + relevel(as.factor(mracef),ref = "white")*as.factor(smokef), data = babies, family = bino
mial)

#change in deviance tests to see if this interaction between smoke and race is useful
anova(babiesreg4, babiesreg1,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c + as.factor(med) + relevel(as.factor(mracef), ref = "white") *
##     as.factor(smokef)
## Model 2: Premature ~ date.c + parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c + as.factor(med) + relevel(as.factor(mracef), ref = "white") +
##     as.factor(smoke)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       847     787.91
## 2       851     793.35 -4  -5.4368   0.2453
```

P-value for the interaction is 0.2453, which isn't significant at all. Therefore, there is no evidence that the association between smoking and pre-term birth differs by mother's race.

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

See the "Explore Interaction" section above. I didn't find any significant interaction effect through trial and error, therefore I keep the baseline model.