Methods and Data Analysis #4

Anna Berman

10/11/2018

Introduction

The following report analyzes a subset of the Child Health and Development Studies, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA. Broadly, our analysis is focused on the relationship between smoking and birth weight. Specifically, our interests are three-fold:

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?
2. Is there any evidence that the association between smoking and pre-term birth differs by mother's race? If so, characterize those differences.
3. Are there other interesting associations with the odds of pre-term birth that are worth mentioning?

Data Overview

The original Child Heath and Development Studies included 15,000 families, however our subset of data includes observations of 1,236 male single births where the baby lived at least 28 days.

Our data is further subset to exclude observations with missing values. Based on the results of our exploratory analysis and model fitting, we removed observations that are missing values for either our outcome or our final predictors. A summary of the remaining dataset is below:

```
# Import dataset to be cleaned
smoke_NA <- read.csv('babiesdata.csv')

# Data cleaning
smoke <- smoke_NA %>%
    # Id and date is not relevent for this analysis
    # Time is not relevent for these observations (all or none)
    # Gestation and birthweight are bivariate predictors of our outcome Premature
    # Remove dht and dwt for too many missing variables
    dplyr::select(-id, -date, -time, -gestation, -bwt.oz, -number, -dht, -dwt) %>%
    # Remove NA observations for variables in our model
    filter(!is.na(Premature),
           !is.na(smoke),
           !is.na(mrace),
           mrace != 10,
           !is.na(mht),
           !is.na(mpregwt),
           !is.na(parity),
           !is.na(inc),
           !is.na(marital),
           !is.na(med),
           med < 6,
           !is.na(mage)) %>%
    # Make smoke a factor
    mutate(smoke = factor(smoke, levels = c('0', '1'))) %>%
```

```
    # Make mother's race a factor
    mutate(mraceF = ifelse(mrace < 6, 'white',
                           ifelse(mrace == 6, 'mexican',
                                  ifelse(mrace == 7, 'black',
                                         ifelse(mrace == 8, 'asian', 'mix'))))) %>%
    mutate(mraceF = factor(mraceF, levels = c('white', 'black', 'mexican',
                                              'asian', 'mix'))) %>%
    # Make mother's race a factor with collapsed baseline with white and mixed
    mutate(mraceF2 = ifelse((mrace < 6 | mrace == 9), 'white or mixed',
                           ifelse(mrace == 6, 'mexican',
                                  ifelse(mrace == 7, 'black',
                                         ifelse(mrace == 8, 'asian', 'error'))))) %>%
    mutate(mraceF2 = factor(mraceF2, levels = c('white or mixed', 'black', 'mexican',
                                              'asian'))) %>%
    # Mean center the numerical predictors except parity
    mutate(mpregwtC = mpregwt - mean(mpregwt),
           mhtC = mht - mean(mht),
           mageC = mage - mean(mage),
           incC = inc - median(inc)) %>%
    # Make med a factor
    mutate(medF = ifelse(med == 0, '< 8th grade',
                         ifelse(med == 1, '8-12 grade',
                                ifelse(med == 2, 'HS only',
                                       ifelse(med == 3, 'HS + trade',
                                              ifelse(med == 4, 'HS + some college',
                                                     ifelse(med ==5, 'college',
                                                            'error'))))))) %>%
    # One copy of the med variable for plotting
    mutate(medF = factor(medF, levels = c('< 8th grade', '8-12 grade', 'HS only',
                                          'HS + trade', 'HS + some college',
                                          'college'))) %>%
    # A second copy of the med variable rebased with HS only
    mutate(medF2 = factor(medF, levels = c('HS only', '< 8th grade', '8-12 grade',
                                          'HS + trade', 'HS + some college',
                                          'college'))) %>%
    # Binned med
    mutate(med_bin = ifelse(med < 2, '< HS',
                            ifelse(med == 2, 'HS only',
                                   ifelse(med == 3 | med == 4, 'HS + trade or some college',
                                          'college')))) %>%
    # One copy of the med_bin variable for plotting
    mutate(med_binF = factor(med_bin, levels = c('< HS', 'HS only',
                                          'HS + trade or some college','college'))) %>%
    # A second copy of the med_bin variable rebased with HS only
    mutate(med_binF2 = factor(med_binF, levels = c('HS only', '< HS',
                                          'HS + trade or some college','college'))) %>%
    # Binned parity
    mutate(parity2 = ifelse(parity < 7, '< 7', '7+')) %>%
    mutate(parity2 = as.factor(parity2)) %>%
    # Binned marital
    mutate(marital2 = ifelse(marital == 1, 'married', 'unmarried')) %>%
    mutate(marital2 = as.factor(marital2)) %>%
    # Income factor
```

```
    mutate(incCF = factor(incC, levels = c(0, -1, -2, -3, 1, 2, 3, 4, 5, 6)))


summary(smoke)
```

```
     parity           mrace            mage            med        
 Min.   : 0.000   Min.   :0.000   Min.   :15.00   Min.   :0.000  
 1st Qu.: 1.000   1st Qu.:0.000   1st Qu.:23.00   1st Qu.:2.000  
 Median : 2.000   Median :2.000   Median :26.00   Median :2.000  
 Mean   : 1.951   Mean   :2.998   Mean   :27.29   Mean   :2.913  
 3rd Qu.: 3.000   3rd Qu.:7.000   3rd Qu.:31.00   3rd Qu.:4.000  
 Max.   :11.000   Max.   :9.000   Max.   :45.00   Max.   :5.000  

      mht            mpregwt          drace             dage       
 Min.   :53.00   Min.   : 87.0   Min.   : 0.000   Min.   :18.00  
 1st Qu.:62.00   1st Qu.:113.0   1st Qu.: 0.000   1st Qu.:25.00  
 Median :64.00   Median :125.0   Median : 3.000   Median :29.00  
 Mean   :64.07   Mean   :128.5   Mean   : 3.156   Mean   :30.21  
 3rd Qu.:66.00   3rd Qu.:140.0   3rd Qu.: 7.000   3rd Qu.:34.00  
 Max.   :72.00   Max.   :220.0   Max.   :10.000   Max.   :62.00  
                                 NA's   :4        NA's   :4      
      ded            marital           inc        smoke      Premature     
 Min.   :0.000   Min.   :0.000   Min.   :0.00   0:463   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.00   1:402   1st Qu.:0.0000  
 Median :4.000   Median :1.000   Median :3.00           Median :0.0000  
 Mean   :3.128   Mean   :1.029   Mean   :3.69           Mean   :0.1861  
 3rd Qu.:5.000   3rd Qu.:1.000   3rd Qu.:5.00           3rd Qu.:0.0000  
 Max.   :7.000   Max.   :5.000   Max.   :9.00           Max.   :1.0000  
 NA's   :7                                                              
     mraceF              mraceF2       mpregwtC            mhtC          
 white  :623   white or mixed:638   Min.   :-41.496   Min.   :-11.07399  
 black  :169   black         :169   1st Qu.:-15.496   1st Qu.: -2.07399  
 mexican: 24   mexican       : 24   Median : -3.496   Median : -0.07399  
 asian  : 34   asian         : 34   Mean   :  0.000   Mean   :  0.00000  
 mix    : 15                       3rd Qu.: 11.504   3rd Qu.:  1.92601  
                                   Max.   : 91.504   Max.   :  7.92601  

     mageC              incC                          medF       
 Min.   :-12.286   Min.   :-3.0000   < 8th grade      :  5  
 1st Qu.: -4.286   1st Qu.:-1.0000   8-12 grade       :130  
 Median : -1.286   Median : 0.0000   HS only          :321  
 Mean   :  0.000   Mean   : 0.6902   HS + trade       : 47  
 3rd Qu.:  3.714   3rd Qu.: 2.0000   HS + some college:203  
 Max.   : 17.714   Max.   : 6.0000   college          :159  

               medF2       med_bin         
 HS only          :321   Length:865         
 < 8th grade      :  5   Class :character   
 8-12 grade       :130   Mode  :character   
 HS + trade       : 47                      
 HS + some college:203                      
 college          :159                      

                     med_binF                       med_binF2
```

```
< HS                          :135   HS only                       :321
HS only                       :321   < HS                          :135
HS + trade or some college:250   HS + trade or some college:250
college                       :159   college                       :159
```

```
parity2        marital2       incCF
< 7:841   married  :846   -2     :151
7+ : 24   unmarried: 19   -1     :145
                          0      :135
                          4      :111
                          1      :105
                          2      : 98
                          (Other):120
```

Mother's Race

Before fitting our model, we altered our Mother's race variable. Given our research question, "Is there any evidence that the association between smoking and pre-term birth differs by mother's race?" fitting a model with interaction effects between smoke and mother's race is critical. However, when these interaction effects were introduced to our model we saw enormous increases in standard errors around the coefficients for mixed race variables. Upon closer examination we see that only 15 mother's in our dataset are mixed race. Furthermore, 100% of mixed race non-smokers had regular term babies and 33.3% of mixed race smokers had premature babies.

Given this distribution, we collapsed our mixed race mother's with the baseline (white mother's). The collapse version of mother's race variable is used throughout our analysis.

Mother's Race (Unaltered)		Non-smoking		Smoking	
	n	Normal Term	Premature	Normal Term	Premature
white	623	0.88	0.12	0.81	0.19
black	169	0.73	0.27	0.73	0.27
mexican	24	0.78	0.22	0.67	0.33
asian	34	0.72	0.28	0.56	0.44
mix	15	1.00	0.00	0.67	0.33

Mother's Race (Altered)		Non-smoking		Smoking	
	n	Normal Term	Premature	Normal Term	Premature
white or mixed	638	0.88	0.12	0.81	0.19
black	169	0.73	0.27	0.73	0.27
mexican	24	0.78	0.22	0.67	0.33
asian	34	0.72	0.28	0.56	0.44

Mother's Education

We also chose to collapse mother's education. Primarily due to low sample sizes, we chose to collapse less than 8th grade education and 8-12th grade education - all levels of education less than completion of high school as well as collapse all education beyond high school without completing college.

Mother's Education (Unaltered)		Non-smoking		Smoking	
	n	Normal Term	Premature	Normal Term	Premature
< 8th grade	5	1.00	0.00	0.00	1.00
8-12 grade	130	0.69	0.31	0.74	0.26
HS only	321	0.85	0.15	0.77	0.23
HS + trade	47	0.63	0.37	0.86	0.14
HS + some college	203	0.91	0.09	0.84	0.16
college	159	0.84	0.16	0.82	0.18

Mother's Education (Altered)		Non-smoking		Smoking	
	n	Normal Term	Premature	Normal Term	Premature
< HS	135	0.71	0.29	0.72	0.28
HS only	321	0.85	0.15	0.77	0.23
HS + trade or some college	250	0.87	0.13	0.84	0.16
college	159	0.84	0.16	0.82	0.18

Parity

As seen in our exploratory analysis, the effects of parity on pre-term birth are prominent when comparing mothers with parity 0-6 and mothers with parity 7+. Therefore we chose to collapse parity into two categories.

```
Warning in cbind(n, round(par_t2, 2), round(par_t3, 2)): number of rows of
result is not a multiple of vector length (arg 1)
```

Parity (Unaltered)		Non-smoking		Smoking	
	n	Normal Term	Premature	Normal Term	Premature
0	209	0.84	0.16	0.82	0.18
1	218	0.81	0.19	0.78	0.22
2	173	0.87	0.13	0.76	0.24
3	119	0.87	0.13	0.82	0.18
4	60	0.79	0.21	0.74	0.26
5	40	0.95	0.05	0.85	0.15
6	22	0.92	0.08	0.70	0.30
7	12	0.50	0.50	1.00	0.00
8	3	1.00	0.00	1.00	0.00
9	5	1.00	0.00	0.67	0.33
11	2	0.00	1.00	0.00	1.00

Parity (Altered)		Non-smoking		Smoking	
	n	Normal Term	Premature	Normal Term	Premature
<7	841	0.85	0.15	0.79	0.21
7+	24	0.57	0.43	0.70	0.30

Marital

As seen in our exploratory analysis, there very few observations representing any marital status other than married. For this reason, we chose to collapse marital status into two categories - married and unmarried.

Marital (Unaltered)		Non-smoking		Smoking	
	n	Normal Term	Premature	Normal Term	Premature
legally seperated	2	0.50	0.50	0.00	0.00
married	846	0.84	0.16	0.79	0.21
divorced	11	1.00	0.00	0.43	0.57
widowed	4	1.00	0.00	0.00	1.00
never married	2	1.00	0.00	1.00	0.00

Marital (Altered)		Non-smoking		Smoking	
	n	Normal Term	Premature	Normal Term	Premature
married	846	0.84	0.16	0.79	0.21
unmarried	19	0.90	0.10	0.44	0.56

Exploratory Analysis

Marginal Plots

Given our research questions, we select pre-term birth (Premature) as our outcome variable and smoking (smoke) as our first predictor variable. Understanding that the relationship between premature birth and smoking might be mediated by other variables, we start examining the relationship between prematurity and other variables that are not also bivariate predictors of premature birth (as are gestational age and birth weight).

```
# CATEGORICAL VARIABLES
# Comparing mean proportion of premature births

# SMOKE
# Does seem to be an association with smoking and pre-term birth
smoke_t <- tapply(smoke$Premature, smoke$smoke, mean)
names(smoke_t) <- c('Non-smoking', 'Smoking')
smoke_t
```

```
Non-smoking     Smoking
  0.1619870   0.2139303
```

```
# MRACE
# Baseline seems to be lower risk than minorities
tapply(smoke$Premature, smoke$mraceF2, mean)
```

```
white or mixed          black        mexican          asian
     0.1551724      0.2662722      0.2500000      0.3235294
```

```
# MED
# Potentially negative association with increased education
tapply(smoke$Premature, smoke$med_binF, mean)
```

```
                   < HS                   HS only
              0.2814815                 0.1900312
HS + trade or some college                   college
              0.1400000                 0.1698113
```

```
# MARITAL
# Seems to be an association between unmarried and pre-term
marital_t <- tapply(smoke$Premature, smoke$marital2, mean)
names(marital_t) <- c('married', 'unmarried')
marital_t
```

```
  married unmarried
0.1832151 0.3157895
```

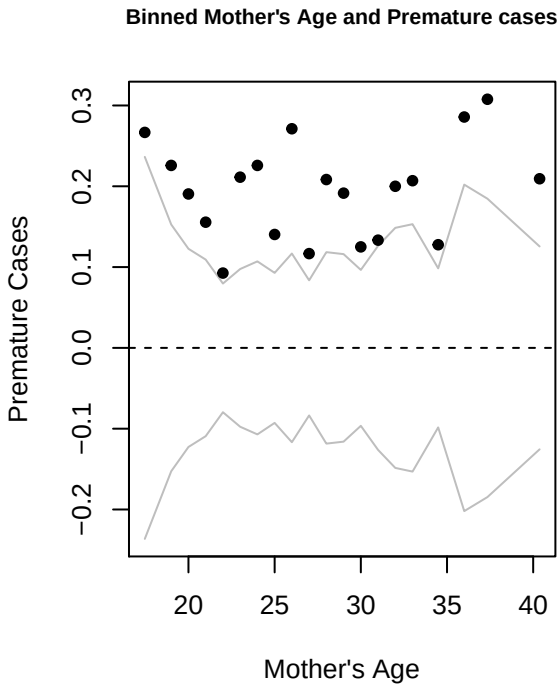
```
# INC
# Unclear effect
inc_t <- tapply(smoke$Premature, smoke$inc, mean)
names(inc_t) <- c('< 2500', '2500 - 4999', '5000 - 7499', '7500 - 9999',
                  '10000 - 12499', '12500 - 14999', '15000 - 17499', '17500 - 19999',
                  '20000 - 22499', '22500+')
inc_t
```

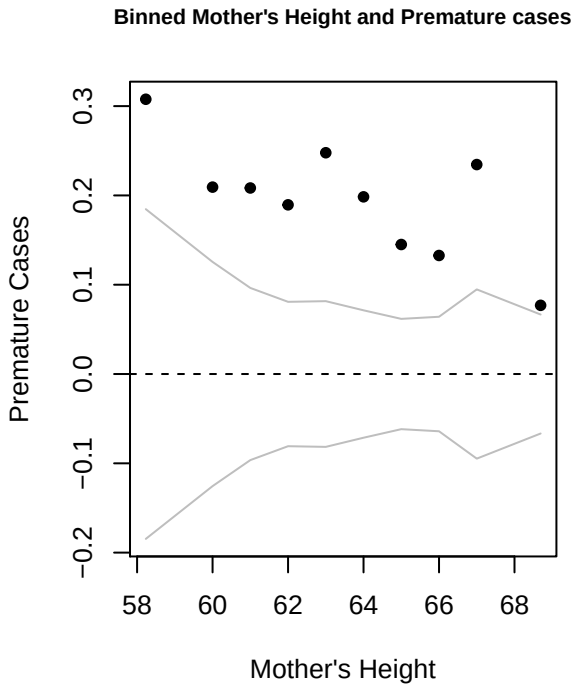
```
       < 2500   2500 - 4999   5000 - 7499   7500 - 9999 10000 - 12499
    0.2692308     0.2052980     0.1586207     0.1703704     0.1904762
12500 - 14999 15000 - 17499 17500 - 19999 20000 - 22499        22500+
    0.2040816     0.1929825     0.1891892     0.0625000     0.1904762
```

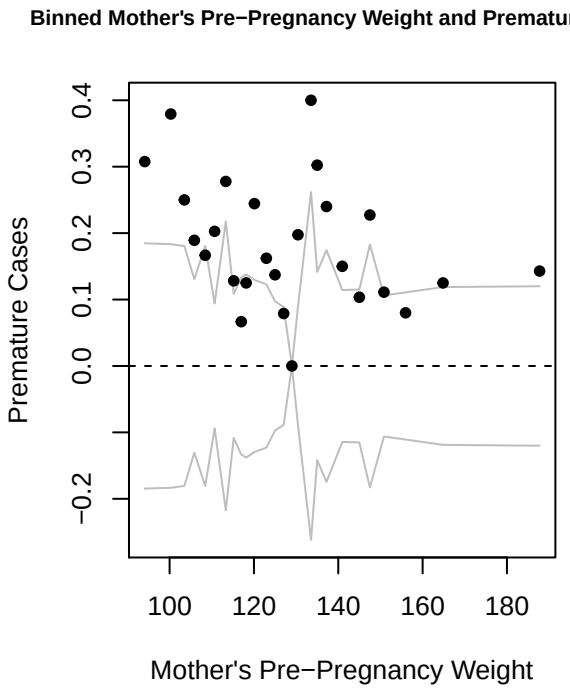
```
# PARITY
# Potentially an association with 7+ parity and premature births
tapply(smoke$Premature, smoke$parity2, mean)
```

```
      < 7        7+
0.1807372 0.3750000
```

```
# CONTINUOUS VARIABLES

par(mfrow = c(2,2))
# MAGE
# Potentially a quadratic effect, but unclear
# (Additional of quadtraic age term did not improve model)
binnedplot(x = smoke$mage, y = smoke$Premature,
           xlab = 'Mother\'s Age', ylab = 'Premature Cases',
           main = 'Binned Mother\'s Age and Premature cases',
           cex.main = .8)

# MHT
# Potentially a negative linear association
binnedplot(x = smoke$mht, y = smoke$Premature,
           xlab = 'Mother\'s Height', ylab = 'Premature Cases',
           main = 'Binned Mother\'s Height and Premature cases',
           cex.main = .8)

# MPREGWT
# Unclear association
binnedplot(x = smoke$mpregwt, y = smoke$Premature,
           xlab = 'Mother\'s Pre-Pregnancy Weight', ylab = 'Premature Cases',
           main = 'Binned Mother\'s Pre-Pregnancy Weight and Premature cases',
           cex.main = .8)
```







Interaction Effects

Thinking specifically about our second research question, "Is there any evidence that the association between smoking and pre-term birth differs by mother's race?", we want to make sure we check for interaction effects between smoking and mother's race. Looking at the results, we see the potential for interaction effects between smoking and mother's race on premature birth.

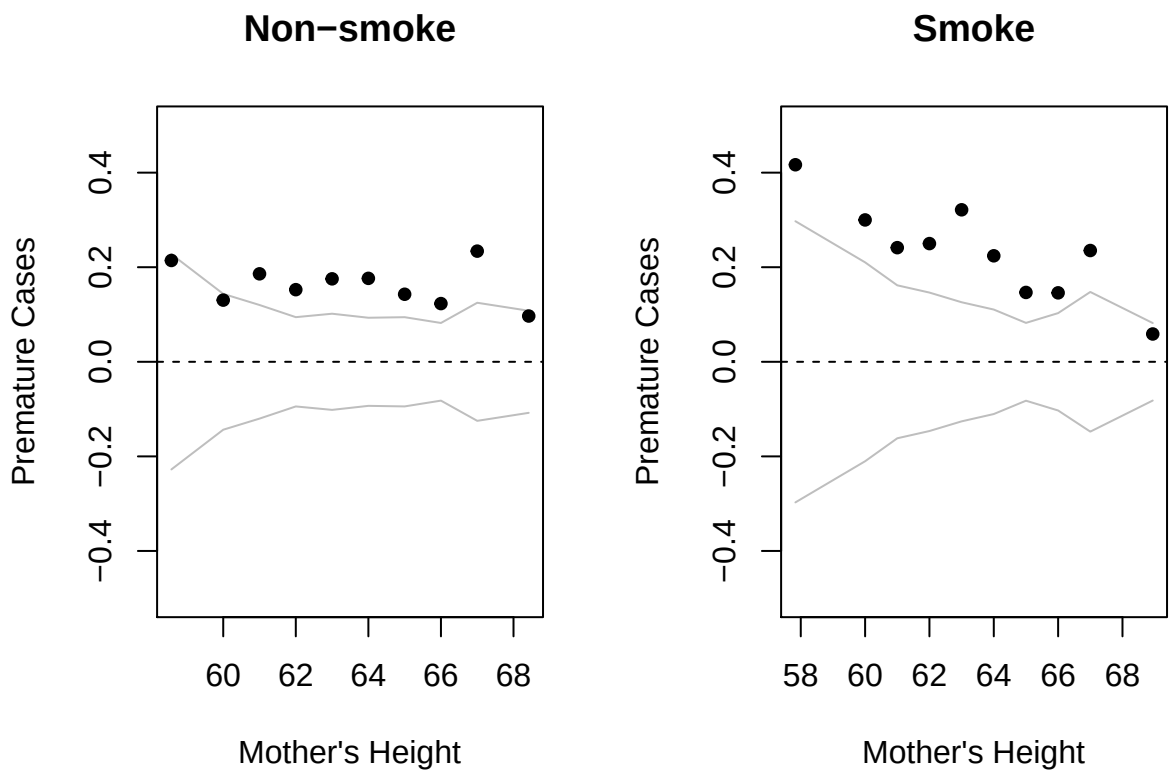
```
# SMOKE AND MRACE
# Potentially interaction effects, specifically different with Asians, but not clear
# Non-smokers
Nonsmokers <- tapply(smoke[smoke$smoke == 0, 'Premature'], smoke[smoke$smoke == 0, 'mraceF2'],
            mean)
# Smokers
Smokers <- tapply(smoke[smoke$smoke == 1, 'Premature'], smoke[smoke$smoke == 1, 'mraceF2'],
            mean)
# Print table
cbind(Nonsmokers, Smokers) %>%
    kable()
```

	Nonsmokers	Smokers
white or mixed	0.1196319	0.1923077
black	0.2659574	0.2666667
mexican	0.2222222	0.3333333
asian	0.2800000	0.4444444

Given that our primary research question revolves around the effects of smoking, we also considered interaction effects between smoking and other predictors. Of note, we did see indication of what could be an interaction effect between smoking and mother's height. Although there is possible evidence for such an interaction, there is little reason to believe that the effects of smoking on pre-term differ between short and tall mothers. Due to the lack of scientific reasoning behind this effect we do not include it in our model. Nevertheless, it should be noted that the addition of this effect to our final model results in a change of deviance with a p-value of 0.06664. Further research should be done to investigate the association between height and the effect of smoking on pre-term birth.

Beyond these interactions however, we did not find evidence for interaction effects between smoking and any other predictors.

```
# SMOKE and MHT
# Seems to be an interaction effect
par(mfrow = c(1,2))
binnedplot(x = smoke[smoke$smoke == 0, 'mht'], y = smoke[smoke$smoke == 0, 'Premature'],
           xlab = 'Mother\'s Height',
           ylab = 'Premature Cases',
           main = 'Non-smoke',
           ylim = c(-.5,.5))
binnedplot(x = smoke[smoke$smoke == 1, 'mht'], y = smoke[smoke$smoke == 1, 'Premature'],
           xlab = 'Mother\'s Height',
           ylab = 'Premature Cases',
           main = 'Smoke',
           ylim = c(-.5,.5))
```



Checking for Multicolinearity

Before running our model, we created a correlation matrix using the numerical variables in our dataset. Most correlations were not concerning, with the most remarkable being a 0.527 correlation between mother's age and parity. However, 0.527 is not high enough to lead us to remove either variable from our list of potential predictors.

Secondly, there could be reason to believe that including both mother's height and weight would introduce effects of mulicollinarity into our model, there is only a 0.46 correlation between mother's height and weight. Therefore we are comfortable including both mother's height and weight in our model.

Model Selection

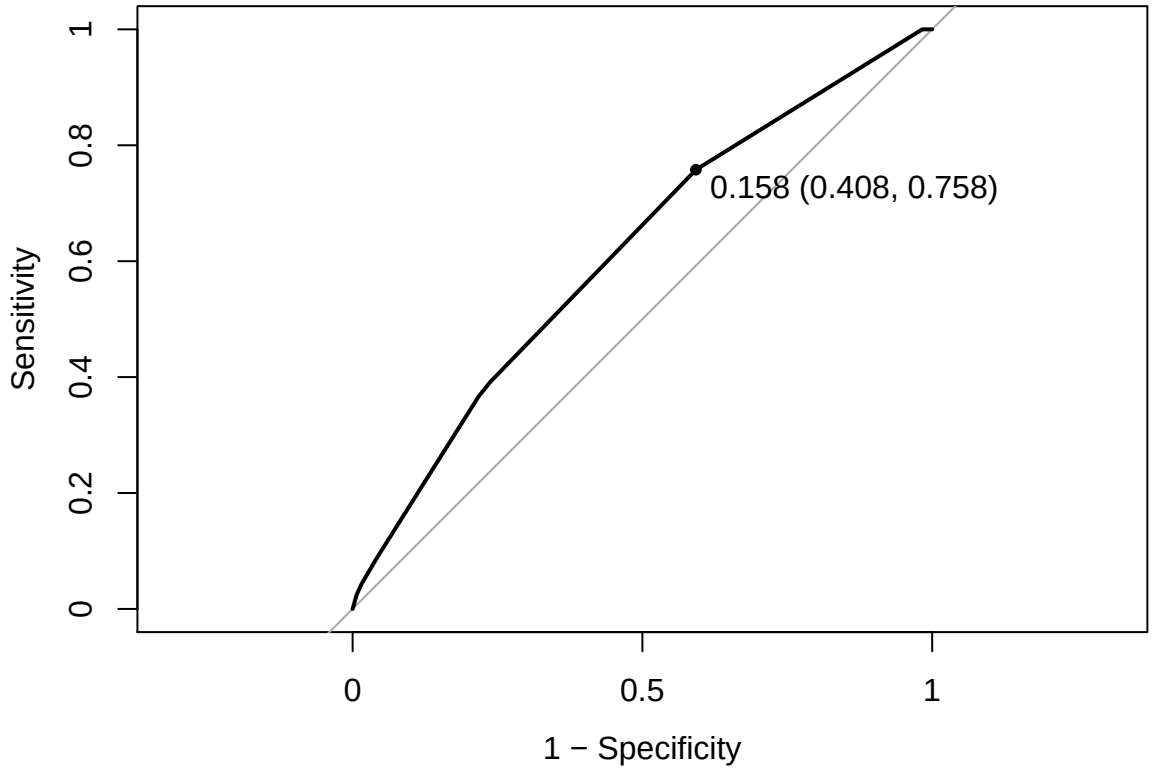
When fitting a model, we work with mean-centered versions of our numerical variables and a median-centered income variable. To make sure we answer our research question, we began fitting our by modeling prematurity on just smoke and race with an interaction effect between the two.

```
# Fit 1
# Smoke only
fit1 <- glm(Premature ~ smoke*mraceF, data = smoke, family = binomial)
summary(fit1)
```

```
Call:
glm(formula = Premature ~ smoke * mraceF, family = binomial,
    data = smoke)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
```

```
-1.084  -0.651  -0.515  -0.515   2.042

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.95321    0.17111 -11.415  < 2e-16 ***
smoke1                0.50929    0.22411   2.272  0.02306 *
mraceFblack           0.93798    0.28943   3.241  0.00119 **
mraceFmexican         0.70045    0.59220   1.183  0.23690
mraceFasian           1.00875    0.47717   2.114  0.03451 *
mraceFmix           -13.61286  420.13711  -0.032  0.97415
smoke1:mraceFblack   -0.50566    0.41581  -1.216  0.22396
smoke1:mraceFmexican  0.05033    1.05908   0.048  0.96210
smoke1:mraceFasian    0.21203    0.83585   0.254  0.79975
smoke1:mraceFmix     14.36364  420.13892   0.034  0.97273
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.37  on 864  degrees of freedom
Residual deviance: 805.38  on 855  degrees of freedom
AIC: 825.38

Number of Fisher Scoring iterations: 14
```

```r
# ROC curve
# Area under the curve: 0.6307
roc1 <- roc(smoke$Premature, fitted(fit1), plot=T, legacy.axes=T, print.thres="best")
```



```
auc(roc1)
```

```
Area under the curve: 0.6139
```

```
# Confusion matrix
threshold <- 0.158
table(smoke$Premature, fit1$fitted > threshold)
```

```
   FALSE TRUE
  0   287  417
  1    39  122
```

This first model has an area under the curve of 0.614. Next, we added in all of the additional terms to our model that we are interested in controling for in our model's interpretation.

```
# Fit 2
# Adding in variables that we want to control for when interpreting findings
# Mother's age, height, and weight
fit2 <- glm(Premature ~ smoke*mraceF2 + mageC + mhtC + mpregwtC + parity2 +
              med_binF2 + incCF + marital2,
            data = smoke, family = binomial)
#summary(fit2)
# P value = 0.08414
anova(fit1, fit2, test= "Chisq")
```

```
Analysis of Deviance Table

Model 1: Premature ~ smoke * mraceF
Model 2: Premature ~ smoke * mraceF2 + mageC + mhtC + mpregwtC + parity2 +
    med_binF2 + incCF + marital2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       855     805.38
2       840     782.38 15       23  0.08414 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Area under the curve: 0.6677
roc2 <- roc(smoke$Premature, fitted(fit2))
auc(roc2)
```

```
Area under the curve: 0.6677
```

```
# Select this as your final model
smoke_fit_final <- fit2
summary(smoke_fit_final)
```

```
Call:
glm(formula = Premature ~ smoke * mraceF2 + mageC + mhtC + mpregwtC +
    parity2 + med_binF2 + incCF + marital2, family = binomial,
    data = smoke)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3466  -0.6787  -0.5406  -0.3949   2.5617

Coefficients:
```

```
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.921111   0.291889  -6.582 4.65e-11
smoke1                            0.522980   0.229942   2.274 0.022942
mraceF2black                      1.098870   0.321987   3.413 0.000643
mraceF2mexican                    0.526857   0.602512   0.874 0.381882
mraceF2asian                      0.775916   0.507847   1.528 0.126549
mageC                             0.006701   0.017675   0.379 0.704617
mhtC                             -0.045853   0.043191  -1.062 0.288401
mpregwtC                         -0.010705   0.005483  -1.952 0.050891
parity27+                         0.671227   0.496438   1.352 0.176348
med_binF2< HS                     0.353726   0.262520   1.347 0.177844
med_binF2HS + trade or some college -0.488178   0.244330  -1.998 0.045713
med_binF2college                 -0.146419   0.278312  -0.526 0.598820
incCF-1                          -0.319036   0.338686  -0.942 0.346202
incCF-2                          -0.150233   0.328452  -0.457 0.647385
incCF-3                           0.414773   0.526913   0.787 0.431179
incCF1                            0.093203   0.352286   0.265 0.791343
incCF2                            0.192122   0.354413   0.542 0.587759
incCF3                            0.073138   0.423363   0.173 0.862845
incCF4                            0.102859   0.350425   0.294 0.769119
incCF5                           -1.239347   1.081641  -1.146 0.251877
incCF6                            0.137253   0.634771   0.216 0.828813
marital2unmarried                 0.568669   0.529929   1.073 0.283225
smoke1:mraceF2black              -0.687109   0.435378  -1.578 0.114522
smoke1:mraceF2mexican            -0.498076   1.116352  -0.446 0.655479
smoke1:mraceF2asian               0.274601   0.856665   0.321 0.748554

(Intercept)                      ***
smoke1                           *
mraceF2black                     ***
mraceF2mexican
mraceF2asian
mageC
mhtC
mpregwtC                         .
parity27+
med_binF2< HS
med_binF2HS + trade or some college *
med_binF2college
incCF-1
incCF-2
incCF-3
incCF1
incCF2
incCF3
incCF4
incCF5
incCF6
marital2unmarried
smoke1:mraceF2black
smoke1:mraceF2mexican
smoke1:mraceF2asian
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.37  on 864  degrees of freedom
Residual deviance: 782.38  on 840  degrees of freedom
AIC: 832.38

Number of Fisher Scoring iterations: 5
```

Through a series of modeling fitting, and change in deviance tests, and comparisons of confusion matrices, we methodically tested alternatives to our model until we were satisfied with the result. Not shown are all the models including variables, manipulations, and interaction effects that did not add explain significant predictive power than a model's without such elements.

Again, although addition of an interaction effect between smoke and mother's height resulted in a change in deviance test with a p-value of .0666, the lack of scientific reasoning ultimately led to a selection of final model without this effect.

Ultimately, we model pre-term birth on smoking, mother's race, age, pre-pregnancy weight, height, parity, education, income and marital status. We include interaction effects of smoking and mother's race, smoking and mother's height, and mother's race and height.

Checking Model Assumptions

Looking at the residuals of our model, average residuals compared to predictors are extremely close to zeo. Based on these results, we are confident that our model fit's our assumptions and we are not missing any major patterns in the data or terms in our model.

```r
# Create raw residuals
rawres <- smoke$Premature - fitted(smoke_fit_final)

par(mfrow = c(2,2))
# MAGEC
binnedplot(x=smoke$mage, y = rawres,
           xlab = "Mother's Age Centered",
           ylab = "Residuals",
           main = "Binned residuals vs. Mother's Age")
# MPREGWT
binnedplot(x=smoke$mpregwt, y = rawres,
           xlab = "Mother's Pre-Pregnancy Weight Centered",
           ylab = "Residuals",
           main = "Binned residuals vs. Mother's Pre-Pregnancy Wieght ")
# MHTC
binnedplot(x=smoke$mht, y = rawres,
           xlab = "Mother's Height Centered",
           ylab = "Residuals",
           main = "Binned residuals vs. Mother's Height")
# MED
tapply(rawres, smoke$med_binF, mean)
```

```
                     < HS                   HS only
             -9.440622e-13             -1.149670e-12
HS + trade or some college                   college
             -4.924382e-13             -2.679931e-12
```

```
# MRACE
tapply(rawres, smoke$mraceF2, mean)
```

```
white or mixed          black        mexican          asian
 -1.278499e-12  -1.361219e-12   1.063890e-16   4.343011e-16
```

```
# PARITY
tapply(rawres, smoke$parity2, mean)
```

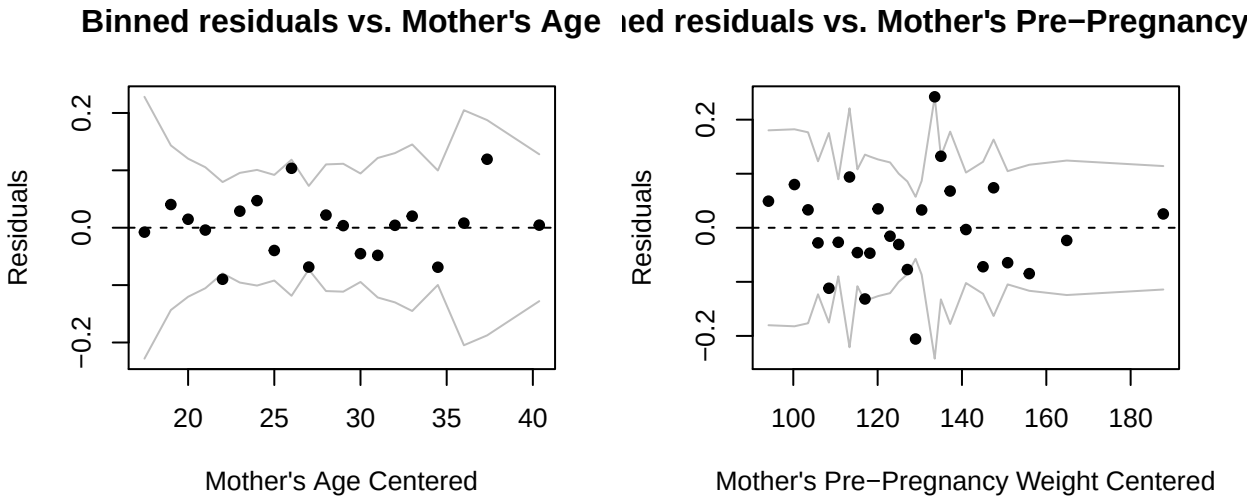
```
          < 7            7+
-1.091826e-12 -5.311905e-12
```

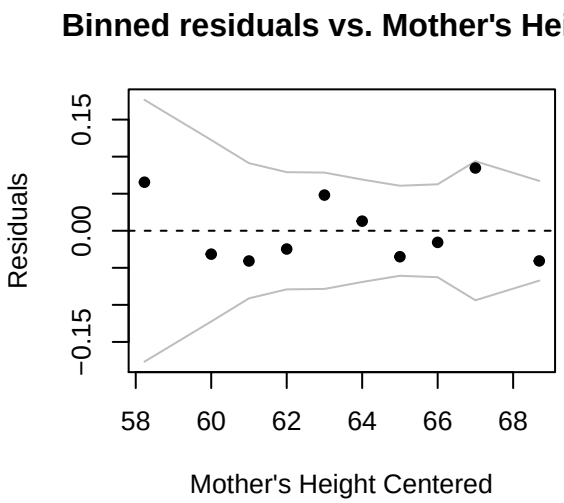
```
# INC
tapply(rawres, smoke$inc, mean)
```

```
            0             1             2             3             4
 4.430842e-17 -1.188400e-16 -1.095413e-16 -3.150352e-16 -2.205784e-17
            5             6             7             8             9
 5.276870e-17 -2.982469e-16  1.874560e-16 -6.535268e-11  1.454767e-17
```

```
# MARTAL
tapply(rawres, smoke$marital2, mean)
```

```
      married     unmarried
-1.236068e-12  1.037096e-16
```





Influential Points

Before we finalize our model, we look for potentially influential points.

```
# Calculate leveage and cooks distance for each observation
leverage <- hatvalues(smoke_fit_final)
cooks <- cooks.distance(smoke_fit_final)

# Append leverage and cooks to our data
smoke_leverage <- smoke %>%
    mutate(leverage, cooks)

# Take a look at the potentially influential points
smoke_leverage %>%
    filter(leverage > .15 | cooks > .03)
```

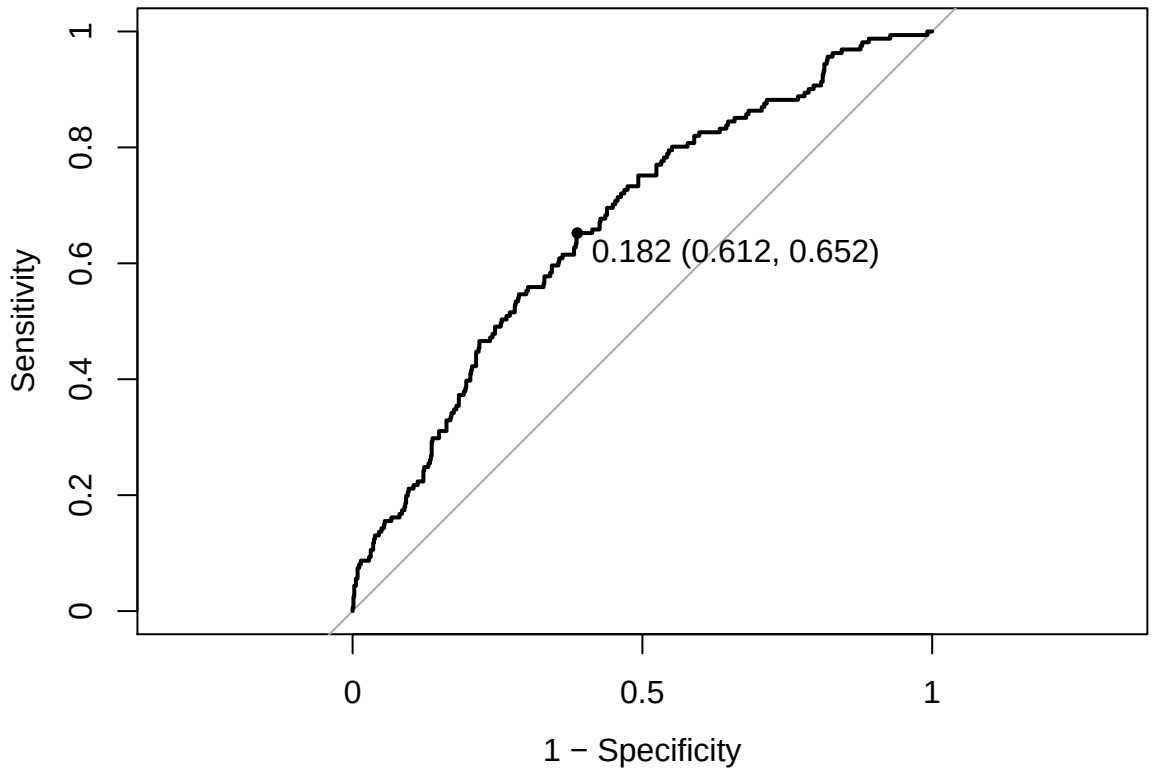
```
  parity mrace mage med mht mpregwt drace dage ded marital inc smoke
1      4     7   32   2  58     130     7   34   1       1   8     0
2     10     0   37   1  65     130     0   40   2       1   8     1
```

```
3     10     6   39   0  53     110     6   41   0       1   3     1
4      1     8   25   5  63     135     8   27   4       1   9     1
5      0     6   18   1  62     110     6   23   2       1   1     1
6      2     6   22   2  62     115     6   24   2       1   1     1
7      4     6   24   1  60     104     6   26   1       1   2     1
8      3     6   29   1  59     100     6   30   1       1   3     1
9      6     6   36   1  63     145     6   38   2       1   1     1
  Premature  mraceF        mraceF2   mpregwtC        mhtC     mageC incC
1         0   black          black   1.504046  -6.0739884  4.714451    5
2         1   white white or mixed   1.504046   0.9260116  9.714451    5
3         1 mexican        mexican -18.495954 -11.0739884 11.714451    0
4         1   asian          asian   6.504046  -1.0739884 -2.285549    6
5         1 mexican        mexican -18.495954  -2.0739884 -9.285549   -2
6         0 mexican        mexican -13.495954  -2.0739884 -5.285549   -2
7         0 mexican        mexican -24.495954  -4.0739884 -3.285549   -1
8         0 mexican        mexican -28.495954  -5.0739884  1.714451    0
9         0 mexican        mexican  16.504046  -1.0739884  8.714451   -2
         medF       medF2 med_bin med_binF med_binF2 parity2 marital2
1     HS only     HS only HS only  HS only   HS only     < 7  married
2  8-12 grade  8-12 grade    < HS     < HS      < HS      7+  married
3 < 8th grade < 8th grade    < HS     < HS      < HS      7+  married
4     college     college college  college   college     < 7  married
5  8-12 grade  8-12 grade    < HS     < HS      < HS     < 7  married
6     HS only     HS only HS only  HS only   HS only     < 7  married
7  8-12 grade  8-12 grade    < HS     < HS      < HS     < 7  married
8  8-12 grade  8-12 grade    < HS     < HS      < HS     < 7  married
9  8-12 grade  8-12 grade    < HS     < HS      < HS     < 7  married
  incCF  leverage       cooks
1     5 0.1504860 0.001424170
2     5 0.1804791 0.053560084
3     0 0.2524696 0.011640566
4     6 0.1892964 0.017058008
5    -2 0.1773250 0.026694877
6    -2 0.1526378 0.002281932
7    -1 0.1825029 0.004407334
8     0 0.2060263 0.008199846
9    -2 0.1591928 0.002619879
```

We can see that the points with the highest leverage or cooks distance in our model are observations from mother's who are primarily Mexican. Unfortunately, our dataset is heavily skewed towards white women, so this is most likely due to a lack of data representing these races. Additionally, several of our highest leverage or cooks distance observations are of very high parity. We are not primarily concerned with modeling pre-term birth at such high parity, thus these are corner cases of our dataset. Finally, we see that many points with the highest leverage or cooks distance have less than high school education. Again, we are not primarily concerned with the effect of less than high school education on the odds of pre-term birth, thus these are corner cases of our dataset. Ultimately, we remain confident that these points do not have a meaningful effect on our model. We select this model as our final model.

Interpretation

```
# ROC curve
roc_final <- roc(smoke$Premature, fitted(smoke_fit_final), plot=T, legacy.axes=T,
                 print.thres="best")
```




```
# Confusion matrix
threshold <- 0.182
matrix <- round(prop.table(table(smoke$Premature, smoke_fit_final$fitted > threshold), 1)*100,1)
table(smoke$Premature, smoke_fit_final$fitted > threshold)
```

```
   FALSE TRUE
  0   433  271
  1    59  102
```

Our final model has an area under the curve of 0.668. Using the suggested threshold of 0.182, our model has a sensitivity of 0.634 and a specificity of 0.385. In other words, our model correctly predicts 63.4% of premature births and 61.5% of normal-term births. This is slightly improved from our first model only inlcuding smoking and mother's race with an area under the curve of 0.614.

```
summary(smoke_fit_final)
```

```
Call:
glm(formula = Premature ~ smoke * mraceF2 + mageC + mhtC + mpregwtC +
    parity2 + med_binF2 + incCF + marital2, family = binomial,
    data = smoke)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3466  -0.6787  -0.5406  -0.3949   2.5617

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.921111   0.291889  -6.582 4.65e-11
smoke1                          0.522980   0.229942   2.274 0.022942
```

```
mraceF2black                           1.098870   0.321987   3.413 0.000643
mraceF2mexican                         0.526857   0.602512   0.874 0.381882
mraceF2asian                           0.775916   0.507847   1.528 0.126549
mageC                                  0.006701   0.017675   0.379 0.704617
mhtC                                  -0.045853   0.043191  -1.062 0.288401
mpregwtC                              -0.010705   0.005483  -1.952 0.050891
parity27+                              0.671227   0.496438   1.352 0.176348
med_binF2< HS                          0.353726   0.262520   1.347 0.177844
med_binF2HS + trade or some college -0.488178   0.244330  -1.998 0.045713
med_binF2college                      -0.146419   0.278312  -0.526 0.598820
incCF-1                               -0.319036   0.338686  -0.942 0.346202
incCF-2                               -0.150233   0.328452  -0.457 0.647385
incCF-3                                0.414773   0.526913   0.787 0.431179
incCF1                                 0.093203   0.352286   0.265 0.791343
incCF2                                 0.192122   0.354413   0.542 0.587759
incCF3                                 0.073138   0.423363   0.173 0.862845
incCF4                                 0.102859   0.350425   0.294 0.769119
incCF5                                -1.239347   1.081641  -1.146 0.251877
incCF6                                 0.137253   0.634771   0.216 0.828813
marital2unmarried                      0.568669   0.529929   1.073 0.283225
smoke1:mraceF2black                   -0.687109   0.435378  -1.578 0.114522
smoke1:mraceF2mexican                 -0.498076   1.116352  -0.446 0.655479
smoke1:mraceF2asian                    0.274601   0.856665   0.321 0.748554

(Intercept)                           ***
smoke1                                *
mraceF2black                          ***
mraceF2mexican
mraceF2asian
mageC
mhtC
mpregwtC                              .
parity27+
med_binF2< HS
med_binF2HS + trade or some college *
med_binF2college
incCF-1
incCF-2
incCF-3
incCF1
incCF2
incCF3
incCF4
incCF5
incCF6
marital2unmarried
smoke1:mraceF2black
smoke1:mraceF2mexican
smoke1:mraceF2asian
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 831.37  on 864  degrees of freedom
Residual deviance: 782.38  on 840  degrees of freedom
AIC: 832.38

Number of Fisher Scoring iterations: 5
```

```r
# Create table summary table
coeff <- data.frame(summary(smoke_fit_final)$coefficients[, c(1,2)],
              confint.default(smoke_fit_final)) %>%
    `colnames<-`(c('Estimate', 'SE', '2.5%', '97.5%'))
```

Intercept: Babies for a white or mixed race, non-smoking, women of average height and weight, with 0-6 previous pregnancies, with a high school only education, an income between $7,500 - $9,999 have an estimated 0.146 odds of pre-term birth (95% CI: 0.083, 0.259).

Smoking: Holding all else constant, for women who have smoked at some point in her life, we estimate the odds of pre-term birth increase by a factor of 1.69 (95% CI: 1.07, 2.65).

Mother's Race Our findings suggest the relationship between odds of pre-term birth and mother's race interacts with smoking in the following way:

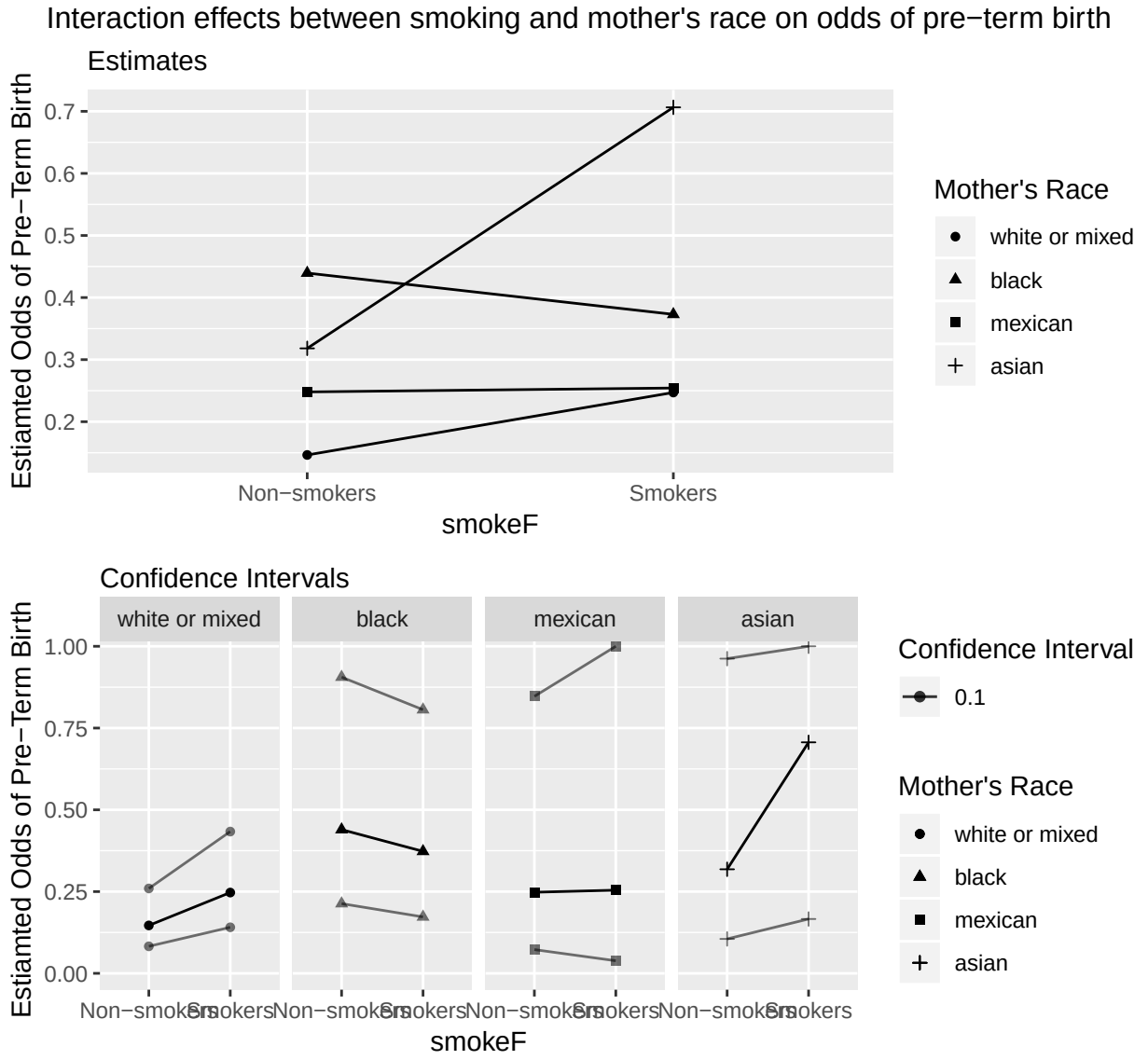
```r
# Rebase our models so we can discuss the impacts of smoking
# for each race individually
# Black
smoke_black <- smoke %>%
    mutate(mraceF2 = factor(mraceF2, levels = c('black', 'white or mixed', 'mexican',
                                                'asian')))

smoke_fit_black <- glm(Premature ~ smoke*mraceF2 + mageC + mhtC + mpregwtC + parity2 +
                med_binF2 + incCF + marital2,
            data = smoke_black, family = binomial)
coeff_black <- data.frame(summary(smoke_fit_black)$coefficients[, c(1,2)],
                          confint.default(smoke_fit_black)) %>%
    `colnames<-`(c('Estimate', 'Std.Error', '2.5%', '97.5%'))

# Mexican
smoke_mex <- smoke %>%
    mutate(mraceF2 = factor(mraceF2, levels = c('mexican', 'white or mixed', 'black',
                                                'asian')))
smoke_fit_mex <- glm(Premature ~ smoke*mraceF2 + mageC + mhtC + mpregwtC + parity2 +
                med_binF2 + incCF + marital2,
            data = smoke_mex, family = binomial)
coeff_mex <- data.frame(summary(smoke_fit_mex)$coefficients[, c(1,2)],
                          confint.default(smoke_fit_mex)) %>%
    `colnames<-`(c('Estimate', 'Std.Error', '2.5%', '97.5%'))

# Asian
smoke_asia <- smoke %>%
    mutate(mraceF2 = factor(mraceF2, levels = c('asian', 'white or mixed', 'black',
                                                'mexican')))
smoke_fit_asia <- glm(Premature ~ smoke*mraceF2 + mageC + mhtC + mpregwtC + parity2 +
                med_binF2 + incCF + marital2,
            data = smoke_asia, family = binomial)
coeff_asia <- data.frame(summary(smoke_fit_asia)$coefficients[, c(1,2)],
                          confint.default(smoke_fit_asia)) %>%
    `colnames<-`(c('Estimate', 'Std.Error', '2.5%', '97.5%'))
```

```r
# Create dummy dataset for charting
newval_race <- data.frame(mhtC = 0,
                      smoke = rep(c(0,1), 4),
                      mpregwtC = 0,
                      mraceF2 = c(rep('white or mixed', 2),
                                  rep('black', 2),
                                  rep('asian', 2),
                                  rep('mexican', 2)),
                      med_binF2 = 'HS only',
                      parity2 = '< 7',
                      mageC = 0,
                      marital2 = 'married',
                      incCF = 0) %>%
    mutate(smoke = as.factor(smoke),
        mraceF2 = factor(mraceF2, levels = c('white or mixed', 'black', 'mexican',
                                             'asian')),
        parity2 = as.factor(parity2),
        marital = as.factor(marital2),
        incCF = as.factor(incCF)) %>%
    mutate(smokeF = ifelse(smoke == 0, 'Non-smokers', 'Smokers'))

# Predict responses
predict <- predict.glm(smoke_fit_final, newval_race, interval = 'response', se.fit = TRUE)

# Create confidence interval
t <- 1.96 ## approx 95% CI
upr <- predict$fit + (t * predict$se.fit)
lwr <- predict$fit - (t * predict$se.fit)
fit <- predict$fit

# Append predictions
newval_race <- newval_race %>%
    mutate(fit = exp(fit),
           lwr = exp(lwr),
           upr = exp(upr))

# Create estimate plot
p1 <- newval_race %>%
    # Cap the odds at 1
    mutate(fit = ifelse(fit < 1, fit, 1),
           upr = ifelse(upr < 1, upr, 1),
           lwr = ifelse(lwr < 1, lwr, 1)) %>%
    ggplot() +
    geom_point(mapping = aes(x = smokeF, y = fit, group = mraceF2,
                                 shape = mraceF2)) +
    geom_line(mapping = aes(x = smokeF, y = fit, group = mraceF2)) +
    labs(subtitle = 'Estimates', shape = 'Mother\'s Race') +
    ylab('Estiamted Odds of Pre-Term Birth')

# Create confidence interval plot
p2 <- newval_race %>%
    # Cap the odds at 1
```

```
    mutate(fit = ifelse(fit < 1, fit, 1),
           upr = ifelse(upr < 1, upr, 1),
           lwr = ifelse(lwr < 1, lwr, 1)) %>%
    ggplot() +
    geom_point(mapping = aes(x = smokeF, y = fit, group = mraceF2,
                                    shape = mraceF2)) +
    geom_line(mapping = aes(x = smokeF, y = fit, group = mraceF2,
                                    shape = mraceF2)) +
    geom_point(mapping = aes(x = smokeF, y = lwr, group = mraceF2,
                                    shape = mraceF2, alpha = .1)) +
    geom_line(mapping = aes(x = smokeF, y = lwr, group = mraceF2,
                                    shape = mraceF2, alpha = .1)) +
    geom_point(mapping = aes(x = smokeF, y = upr, group = mraceF2,
                                    shape = mraceF2, alpha = .1)) +
    geom_line(mapping = aes(x = smokeF, y = upr, group = mraceF2,
                                    shape = mraceF2, alpha = .1)) +
    facet_grid(. ~ mraceF2) +
    labs(subtitle = 'Confidence Intervals', shape = 'Mother\'s Race',
         alpha = 'Confidence Interval') +
    ylab('Estiamted Odds of Pre-Term Birth') +
    ylim(c(0,1))
```

```
Warning: Ignoring unknown aesthetics: shape

Warning: Ignoring unknown aesthetics: shape

Warning: Ignoring unknown aesthetics: shape
```

```
grid.arrange(p1, p2,
             top = 'Interaction effects between smoking and mother\'s race on odds of pre-term birth')
```



Mother's Pre-pregnancy Weight: Holding all else constant, for each additional 10 pounds mothers weighed before pregnancy, we estimate the odds of pre-term birth increase decrease by a factor of 0.9 (95% CI: 0.81, 1). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect of mother's weight on odds of pre-term birth.

Mother's Age: Holding all else constant, for each 5 additional years mothers age we estimate the odds of pre-term birth increase by a factor of 1.03 (95% CI: 0.87, 1.23). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect of mother's age on pre-term birth.

Mother's Height: Holding all else constant, for each additional inch in mothers height we estimate the odds of pre-term birth decrease by a factor of 0.96 (95% CI: 0.88, 1.04). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect of mother's age on pre-term birth.

Parity: Holding all else constant, if a mothers has had 7 or more previous pregnancies, we estimate the odds of pre-term birth increase by a factor of 1.96 (95% CI: 0.74, 5.18). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect of parity on pre-term birth.

Marital: Holding all else constant, if a mothers is unmarried, we estimate the odds of pre-term birth increase by a factor of 1.77 (95% CI: 0.63, 4.99). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect of parity on pre-term birth.

Education Holding all else constant, for a woman with:

- Less than a high school education, we estimate the odds of pre-term birth increase by 1.42 (95% CI: 0.85, 2.38). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect.
- Education including either trade school or some college, we estimate the odds of pre-term birth decrease by 0.61 (95% CI: 0.38, 0.99).
- College education, we estimate the odds of pre-term birth decrease by 0.86 (95% CI: 0.5, 1.49). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect.

Conclusion

Thinking back to our original research questions. In general, our findings suggest, that smoking increases a mother's odds of pre-term birth. However, our findings also suggest that the effect of smoking on odds of pre-term birth differ by race. Specifically,

- White or mixed race women who smoke are estimated to have the odds of pre-term birth increased by a factor of 1.69 (95% CI: 1.07, 2.65) compared to White women who don't smoke.
- Black women who smoke are estimated to have the odds of pre-term birth decreased by a factor of 0.85 (95% CI: 0.41, 1.75) compared to Black women who don't smoke. Given that this confidence interval includes 1, we are not confident that there is a meaningful effect.
- Asian women who smoke are estimated to have the odds of pre-term birth increased by a factor of 2.22 (95% CI: 0.44, 11.19) compared to Asian women who don't smoke. Given that this confidence interval includes 1, we are not confident that there is a meaningful effect.
- Mexican women who smoke are estimated to have the odds of pre-term birth decreased by a factor of 1.03 (95% CI: 0.12, 8.74) compared to Mexican women who don't smoke. Given that this confidence interval includes 0, we are not confident that there is a meaningful difference.

The most unique of these cases are found in Black women, for which our estimates suggest that smoking might actually have be associated with decreased odds of pre-term birth. However, it should be noted that Black women only accounted for 19.5% of our dataset respectively (169. Furthermore, our confidence intervals for this coefficient included 1, therefore we are not confident that this relationship is significant. Additional research is needed to understand the relationship between smoking and birth weight in Mexican women.

Interestingly, our findings also suggest, in Mexican women, smoking has almost no impact on odds of pre-term birth. Again, it should be noted that Mexican women only accounted for 2.8% of our dataset respectively (24. Such low sample sizes contributed to large confidence intervals.

Finally, our findings also suggest that smoking may have heavier effects on birth weight in Asian and compared to White or Multiracial women. In other words, smoking may be associated with higher increased odds of pre-term birth in Asian women compared to White or Multiracial women. Again, it should be noted that Asian women only accounted for 3.9% of our dataset (34 observations. Therefore additional research is needed to understand the strength of the relationship between smoking and birth weight in a diverse population of women.

Beyond smoking and race there are very few variables that appear to have a significant impact on the odds of pre-term birth. There are several factors of variables may have a "statistically significant"" effect on the odds of pre-term birth, such as having an education including trade school or some college for example. However, it's not clear whether these are reflective of true associations, or merely an effect of limited sample size.

Limitations

Our final model has an area under the curve of 0.668. Using the suggested threshold of 0.182, our model has a sensitivity of 0.634 and a specificity of 0.385. In other words, our model correctly identifies 63.4% of true premature births and 61.5% of true normal-term births. It seems that we are missing variables in our model that would explain additional variation in pre-term birth cases, therefore more research is needed to fully understand the relationship between smoking and birth weight and the mediating variables in this relationship.

Additionally, the data included in this study was heavily weighted towards White mothers. Our research suggests that the effect of smoking on birth weight may differ by race, but to be truly confident in these findings, further research should be done on the effect of smoking on birth weights in minorities. Beyond diversification in race, our dataset was also lacking in unmarried mothers and mothers with educations other that high school only and college. To fully understand the effects of these variables, additional data and research is needed.

Additionally, this data was collected as an observational study. To fully understand the causal nature of the relationship between smoking and birth weight, a randomized control trial is needed.