Methods and Data Analysis #3

Anna Berman

9/27/2018

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 give birth to babies with lower weights than mothers who do not smoke? What is a likely range for the difference in birth weights for smokers and non-smokers?
2. Is there any evidence that the association between smoking and birth weight differs by mother's race? If so, what characterizes those differences.
3. Are there other interesting associations between smoking and birth weight 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 subsetted 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 fo the remaining dataset is below:

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

# Data cleaning
smoke <- smoke_NA %>%
    # Time is not relevent for these observations (all or none)
    # Gestation and premature are bivariate predictors of our outcome birth weight
    # Remove dht and dwt for too many missing variables
    select(-time, -gestation, -Premature, -number, -dht, -dwt) %>%
    # Remove NA observations for variables in our model
    filter(!is.na(bwt.oz),
           !is.na(smoke),
           !is.na(mrace),
           mrace != 10,
           !is.na(mht),
           !is.na(mpregwt),
           !is.na(parity),
           !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'))) %>%
    # Mean center the numerical predictors except parity
    mutate(mpregwtC = mpregwt - mean(mpregwt),
           mhtC = mht - mean(mht),
           mageC = mage - mean(mage)) %>%
    # 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')))
summary(smoke)
```

```
       id            date          bwt.oz          parity
 Min.   :  15   Min.   :1350   Min.   : 55.0   Min.   : 0.000
 1st Qu.:5484   1st Qu.:1444   1st Qu.:107.0   1st Qu.: 0.000
 Median :6758   Median :1540   Median :119.0   Median : 1.000
 Mean   :6072   Mean   :1537   Mean   :118.6   Mean   : 1.952
 3rd Qu.:7618   3rd Qu.:1627   3rd Qu.:130.0   3rd Qu.: 3.000
 Max.   :9263   Max.   :1714   Max.   :176.0   Max.   :13.000

     mrace            mage            med             mht
 Min.   :0.000   Min.   :15.00   Min.   :0.000   Min.   :53.00
 1st Qu.:0.000   1st Qu.:23.00   1st Qu.:2.000   1st Qu.:62.00
 Median :3.000   Median :26.00   Median :2.000   Median :64.00
 Mean   :3.101   Mean   :27.16   Mean   :2.855   Mean   :64.05
 3rd Qu.:7.000   3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:66.00
 Max.   :9.000   Max.   :45.00   Max.   :5.000   Max.   :72.00

    mpregwt          drace             dage            ded
 Min.   : 87.0   Min.   : 0.000   Min.   :18.00   Min.   :0.000
 1st Qu.:113.0   1st Qu.: 0.000   1st Qu.:25.00   1st Qu.:2.000
 Median :125.0   Median : 3.000   Median :29.00   Median :3.000
 Mean   :128.6   Mean   : 3.257   Mean   :30.11   Mean   :3.078
 3rd Qu.:139.0   3rd Qu.: 7.000   3rd Qu.:34.00   3rd Qu.:5.000
 Max.   :250.0   Max.   :10.000   Max.   :62.00   Max.   :7.000
                 NA's   :4        NA's   :4       NA's   :9
    marital           inc        smoke       mraceF       mpregwtC
 Min.   :0.000   Min.   :0.000   0:519   white  :689   Min.   :-41.588
 1st Qu.:1.000   1st Qu.:2.000   1:455   black  :198   1st Qu.:-15.588
 Median :1.000   Median :3.000           mexican: 29   Median : -3.588
```

```
 Mean   :1.037   Mean   :3.688           asian  : 37   Mean   :  0.000
 3rd Qu.:1.000   3rd Qu.:5.000           mix    : 21   3rd Qu.: 10.412
 Max.   :5.000   Max.   :9.000                          Max.   :121.412
                 NA's   :100
     mhtC              mageC                        medF
 Min.   :-11.0462   Min.   :-12.16   < 8th grade      : 14
 1st Qu.: -2.0462   1st Qu.: -4.16   8-12 grade       :149
 Median : -0.0462   Median : -1.16   HS only          :364
 Mean   :  0.0000   Mean   :  0.00   HS + trade       : 51
 3rd Qu.:  1.9538   3rd Qu.:  3.84   HS + some college:229
 Max.   :  7.9538   Max.   : 17.84   college          :167

              medF2
 HS only          :364
 < 8th grade      : 14
 8-12 grade       :149
 HS + trade       : 51
 HS + some college:229
 college          :167
```

Exploratory Analysis

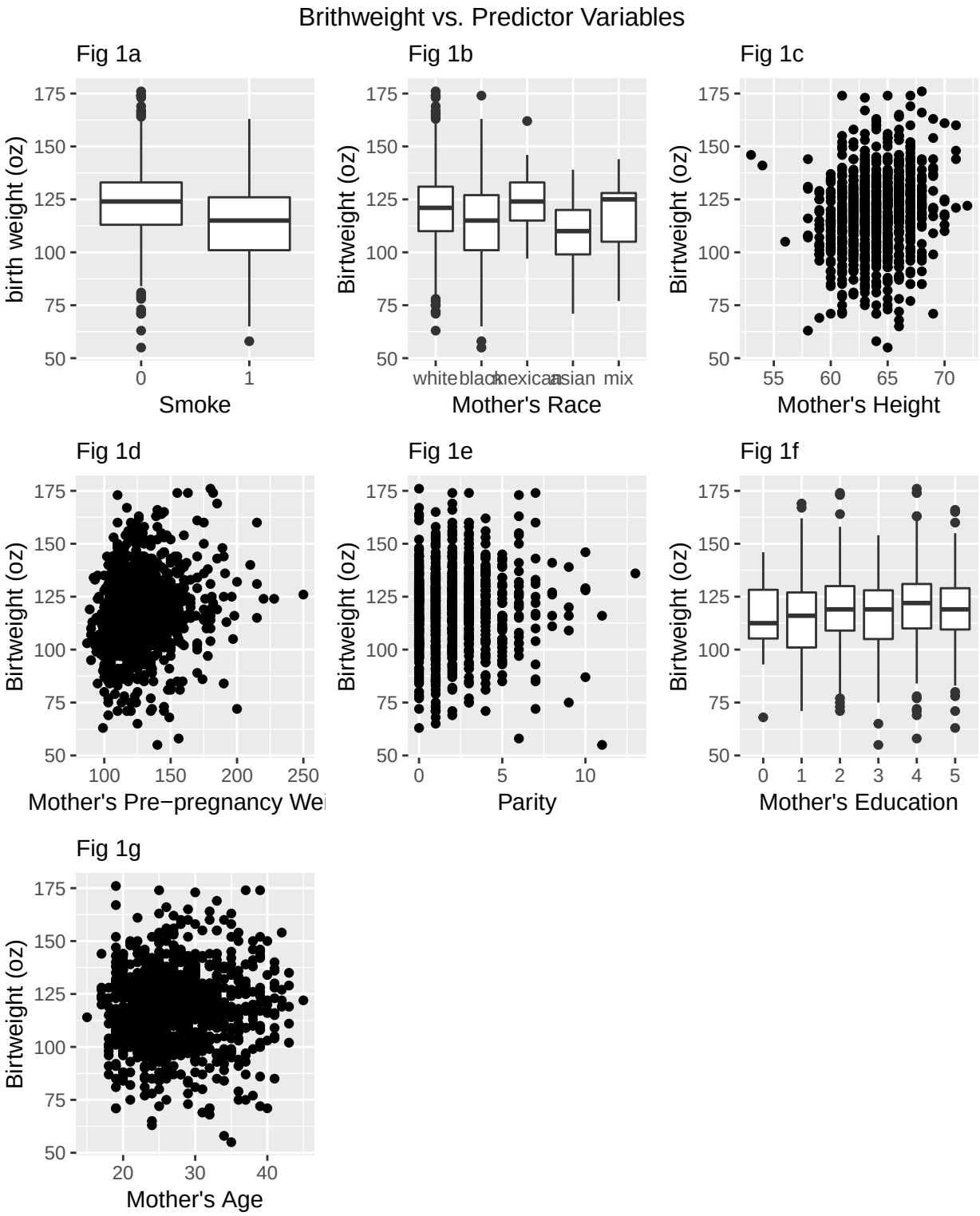
Marginal Plots

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

After examining each relationship through marginal plots, we select the variables in Figure 1 as potentially relevant predictors in our model. We did not select any father variables such as father's race, age, or education due to potential issues with colinearities. We did not select martial status, income, or date because not only did they not seem to have a strong relationship with birth weight and also did not logically seem like they would have an effect on birth weight. Some variables included below such as mother's education and mother's age do not immediately appear to have a large effect on birth weight, but were shown to have larger effects when other variables are controlled for.

```
# SMOKE
# birth weight by smoke
e1 <- ggplot(data = smoke) +
    geom_boxplot(mapping = aes(x = smoke, y = bwt.oz)) +
    xlab('Smoke') +
    ylab('birth weight (oz)') +
    labs(subtitle = "Fig 1a")

# MRACE
# birth weight by mother's race
e2 <- ggplot(data = smoke) +
    geom_boxplot(mapping = aes(x = mraceF, y = bwt.oz)) +
    xlab('Mother\'s Race') +
    ylab('Birtweight (oz)') +
    labs(subtitle = "Fig 1b")
```

```r
# MHT
# birth weight by mother's height
e3 <- ggplot(data = smoke) +
    geom_point(mapping = aes(x = mht, y = bwt.oz)) +
    xlab('Mother\'s Height') +
    ylab('Birtweight (oz)') +
    labs(subtitle = "Fig 1c")

# MPREGWT
# birth weight by mother's pregnant weight
e4 <- ggplot(data = smoke) +
    geom_point(mapping = aes(x = mpregwt, y = bwt.oz))+
    xlab('Mother\'s Pre-pregnancy Weight') +
    ylab('Birtweight (oz)') +
    labs(subtitle = "Fig 1d")

# PARITY
# birth weight by parity
e5 <- ggplot(data = smoke) +
    geom_point(mapping = aes(x = parity, y = bwt.oz)) +
    xlab('Parity') +
    ylab('Birtweight (oz)') +
    labs(subtitle = "Fig 1e")

# MED
# Birthwieght by mother's education
e6 <- ggplot(data = smoke) +
    geom_boxplot(mapping = aes(x = as.factor(med), y = bwt.oz)) +
    xlab('Mother\'s Education') +
    ylab('Birtweight (oz)') +
    labs(subtitle = "Fig 1f")

# MAGE
# Birthwieght by mother's age
e7 <- ggplot(data = smoke) +
    geom_point(mapping = aes(x = mage, y = bwt.oz)) +
    xlab('Mother\'s Age') +
    ylab('Birtweight (oz)') +
    labs(subtitle = "Fig 1g")

grid.arrange(e1, e2, e3, e4, e5, e6, e7,
             top = 'Brithweight vs. Predictor Variables')
```

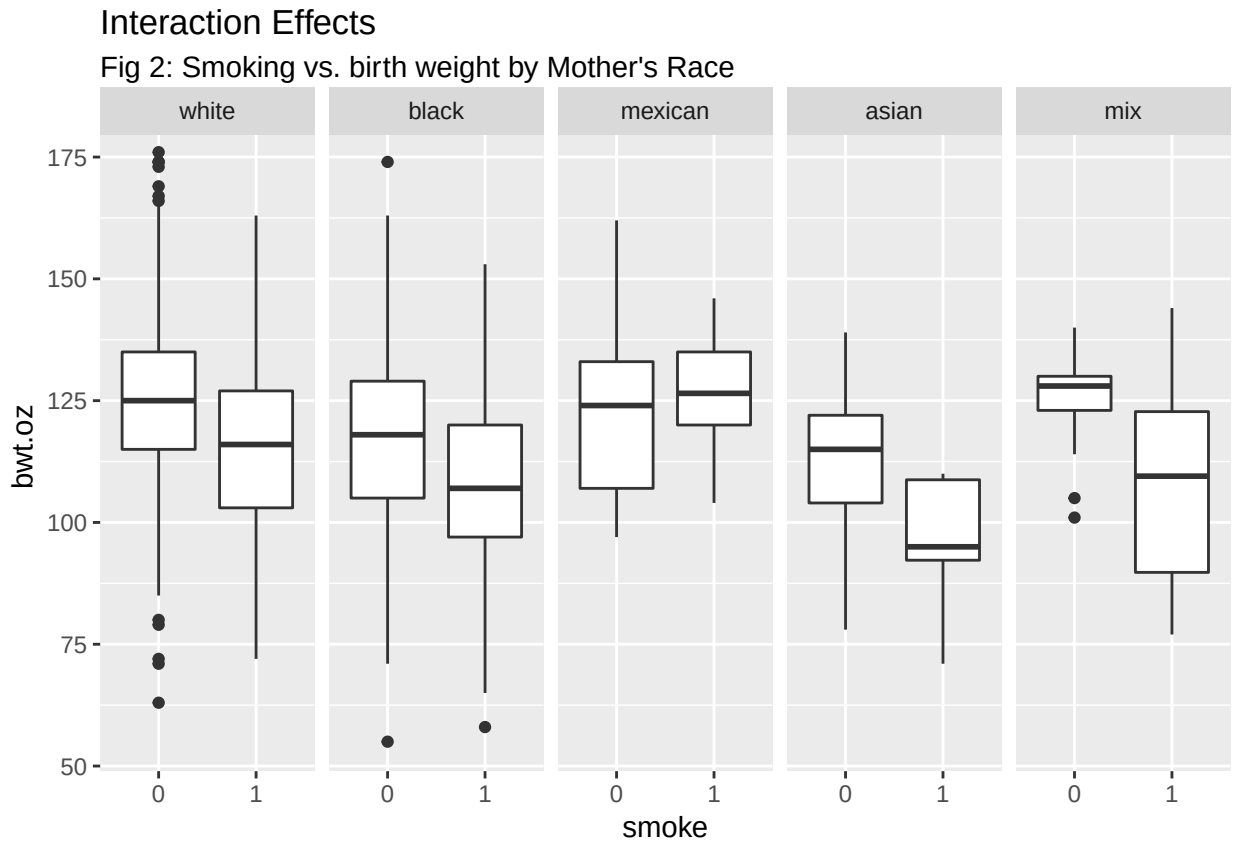


Interaction Effects

Thinking specifically about our second research question, "Is there any evidence that the association between smoking and birth weight 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 in Figure 2a, we see the potential for interaction effects between smoking and mother's race on birth weight. We also considered interaction effects for other categorical predictors, but did not see clear indication of such effects.

```
# INTERACTION EFFECTS
# MRACE
ggplot(data = smoke) +
    geom_boxplot(mapping = aes(x = smoke, y = bwt.oz)) +
    facet_grid(. ~ mraceF) +
    ggtitle('Interaction Effects') +
    labs(subtitle = "Fig 2: Smoking vs. birth weight by Mother's Race")
```



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 0.529 correlation between mother's age and parity. However, 0.529 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 0.443 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. We began fitting our model with just modeling birth weight on smoke.

```r
fit1 <- lm(bwt.oz ~ smoke, data = smoke)
summary(fit1)
```

```
Call:
lm(formula = bwt.oz ~ smoke, data = smoke)

Residuals:
    Min      1Q  Median      3Q     Max
-67.817 -10.895   1.105  11.105  53.183

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 122.8170     0.7779  157.89  < 2e-16 ***
smoke1       -8.9225     1.1381   -7.84 1.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.72 on 972 degrees of freedom
Multiple R-squared:  0.05947,   Adjusted R-squared:  0.0585
F-statistic: 61.46 on 1 and 972 DF,  p-value: 1.183e-14
```

This first model only explains 5.95% the variation in birth weight so we continue to add additional predictors to our model. Through a series of modeling fitting and nested F tests we methodically add additional items to our model until we are satisfied with the result. (Not shown are all the models including variables, manipulations, and interaction effects that did not add explain significantly more variance in btw.oz than a model's without such elements.)

```r
# FITTING A MODEL
# Just smoking compared to what we thinks makes sense logically
#   and what makes sense for interpretation -
#   Mother's race, age, weight, and height
# p value: 2.2e-16
fit1 <- lm(bwt.oz ~ smoke, data = smoke)
fit2 <- lm(bwt.oz ~ smoke + mraceF + mageC + mpregwtC + mhtC, data = smoke)
#summary(fit2)
anova(fit1, fit2)
```

```
Analysis of Variance Table

Model 1: bwt.oz ~ smoke
Model 2: bwt.oz ~ smoke + mraceF + mageC + mpregwtC + mhtC
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    972 305255
2    965 274568  7     30687 15.407 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Adding in parity
# p value: 0.02314
fit1 <- lm(bwt.oz ~ smoke + mraceF + mageC + mpregwtC + mhtC, data = smoke)
fit2 <- lm(bwt.oz ~ smoke + mraceF + mageC + mpregwtC + mhtC + parity,
           data = smoke)
#summary(fit2)
anova(fit1, fit2)
```

```
Analysis of Variance Table

Model 1: bwt.oz ~ smoke + mraceF + mageC + mpregwtC + mhtC
Model 2: bwt.oz ~ smoke + mraceF + mageC + mpregwtC + mhtC + parity
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    965 274568
2    964 273102  1      1466 5.1748 0.02314 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Adding interaction effects of race
# p value: 0.03125
fit1 <- lm(bwt.oz ~ smoke + mraceF + mageC + mpregwtC + mhtC + parity,
           data = smoke)
fit2 <- lm(bwt.oz ~ smoke*mraceF + mraceF + mageC + mpregwtC + mhtC + parity,
           data = smoke)
#summary(fit2)
anova(fit1, fit2)
```

```
Analysis of Variance Table

Model 1: bwt.oz ~ smoke + mraceF + mageC + mpregwtC + mhtC + parity
Model 2: bwt.oz ~ smoke * mraceF + mraceF + mageC + mpregwtC + mhtC +
    parity
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    964 273102
2    960 270102  4    3000.1 2.6657 0.03125 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Adding in MED
# p value: 0.08127
fit1 <- lm(bwt.oz ~ smoke*mraceF + mraceF + mageC + mpregwtC + mhtC + parity,
           data = smoke)
fit2 <- lm(bwt.oz ~ smoke*mraceF + mraceF + mageC + mpregwtC + mhtC + parity +
               medF2, data = smoke)
#summary(fit2)
anova(fit1, fit2)
```

```
Analysis of Variance Table

Model 1: bwt.oz ~ smoke * mraceF + mraceF + mageC + mpregwtC + mhtC +
    parity
Model 2: bwt.oz ~ smoke * mraceF + mraceF + mageC + mpregwtC + mhtC +
    parity + medF2
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    960 270102
2    955 267350  5    2751.8 1.9659 0.08127 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Select this as your final model
smoke_fit_final <- lm(bwt.oz ~ smoke*mraceF + mraceF + mageC + mpregwtC +
                          mhtC + parity + medF2, data = smoke)
summary(smoke_fit_final)
```

```
Call:
lm(formula = bwt.oz ~ smoke * mraceF + mraceF + mageC + mpregwtC +
    mhtC + parity + medF2, data = smoke)

Residuals:
   Min     1Q Median     3Q    Max
-67.64 -10.01  -0.56  10.29  50.48

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             124.40243    1.30941  95.006  < 2e-16 ***
smoke1                   -9.50320    1.30123  -7.303 5.92e-13 ***
mraceFblack             -11.43206    1.92027  -5.953 3.69e-09 ***
mraceFmexican             0.14556    3.84480   0.038 0.969809
mraceFasian              -5.57582    3.45002  -1.616 0.106388
mraceFmix                -0.15606    4.75307  -0.033 0.973814
mageC                    -0.09102    0.11912  -0.764 0.444988
mpregwtC                  0.11568    0.02998   3.859 0.000122 ***
mhtC                      0.98116    0.25207   3.892 0.000106 ***
parity                    0.95854    0.36017   2.661 0.007913 **
medF2< 8th grade         -7.72820    4.78226  -1.616 0.106421
medF28-12 grade          -3.73728    1.72824  -2.162 0.030829 *
medF2HS + trade          -2.41855    2.53219  -0.955 0.339758
medF2HS + some college    0.82463    1.44255   0.572 0.567695
medF2college             -1.64162    1.64855  -0.996 0.319601
smoke1:mraceFblack        2.98799    2.75186   1.086 0.277838
smoke1:mraceFmexican     18.54938    7.15273   2.593 0.009651 **
smoke1:mraceFasian       -9.12004    6.33960  -1.439 0.150597
smoke1:mraceFmix         -8.84448    7.65275  -1.156 0.248083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.73 on 955 degrees of freedom
Multiple R-squared:  0.1763,    Adjusted R-squared:  0.1607
F-statistic: 11.35 on 18 and 955 DF,  p-value: < 2.2e-16
```

Ultimately, we model birth weight on smoking, mother's race, age, pre-pregnancy weight, height parity, and education. We also include interaction effects between smoking and race in our model.

Checking Model Assumptions

Looking at the residuals of our model, they are normal distributed and have constant variance. Based on these results, we are confident that our model fit's our assumptions.

```
# RESIDUALS
r1 <- ggplot() +
    geom_qq(mapping = aes(sample = smoke_fit_final$residuals)) +
    labs(subtitle = 'Fig 3a: qqnorm') +
    ylab('Residuals')
r2 <- ggplot() +
    geom_boxplot(mapping = aes(x = smoke$smoke, y = smoke_fit_final$residuals)) +
    geom_hline(yintercept = 0) +
    xlab('smoke') +
```

```
    ylab('Residuals') +
    labs(subtitle = 'Fig 3b')
r3 <- ggplot() +
    geom_boxplot(mapping = aes(x = smoke$mraceF, y = smoke_fit_final$residuals)) +
    geom_hline(yintercept = 0) +
    xlab('mraceF') +
    ylab('Residuals') +
    labs(subtitle = 'Fig 3c')
r4 <- ggplot() +
    geom_point(mapping = aes(x = smoke$mpregwt, y = smoke_fit_final$residuals)) +
    geom_hline(yintercept = 0) +
    xlab('mpregwt') +
    ylab('Residuals') +
    labs(subtitle = 'Fig 3d')
r5 <- ggplot() +
    geom_point(mapping = aes(x = smoke$mht, y = smoke_fit_final$residuals)) +
    geom_hline(yintercept = 0) +
    xlab('mht') +
    ylab('Residuals') +
    labs(subtitle = 'Fig 3e')
r6 <- ggplot() +
    geom_point(mapping = aes(x = smoke$parity, y = smoke_fit_final$residuals)) +
    geom_hline(yintercept = 0) +
    xlab('parity') +
    ylab('Residuals') +
    labs(subtitle = 'Fig 3f')
r7 <- ggplot() +
    geom_point(mapping = aes(x = smoke$mage, y = smoke_fit_final$residuals)) +
    geom_hline(yintercept = 0) +
    xlab('mage') +
    ylab('Residuals') +
    labs(subtitle = 'Fig 3g')
r8 <- ggplot() +
    geom_point(mapping = aes(x = smoke$medF, y = smoke_fit_final$residuals)) +
    geom_hline(yintercept = 0) +
    xlab('mage') +
    ylab('Residuals') +
    labs(subtitle = 'Fig 3h')

grid.arrange(r1, r2, r3, r4, r5, r6, r7, r8,
             top = 'Residual Plots')
```

Residual Plots

Fig 3a: qqnorm

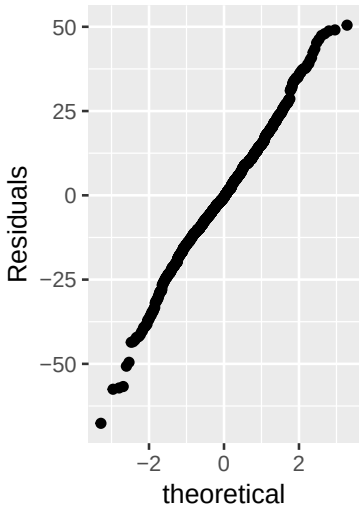


Fig 3b

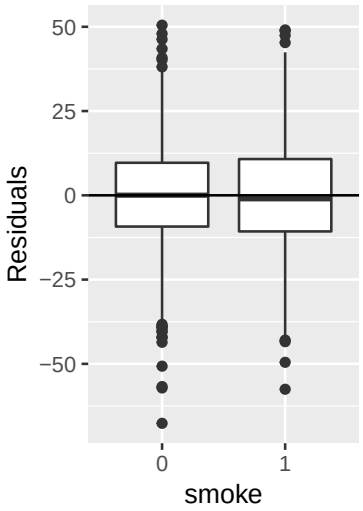


Fig 3c

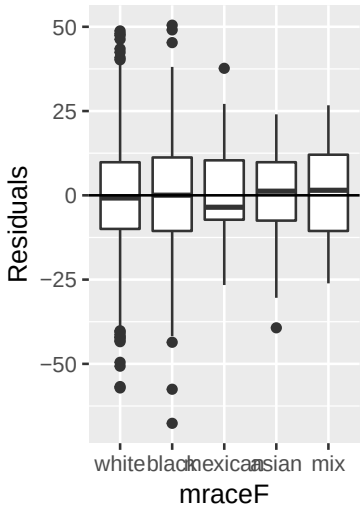


Fig 3d

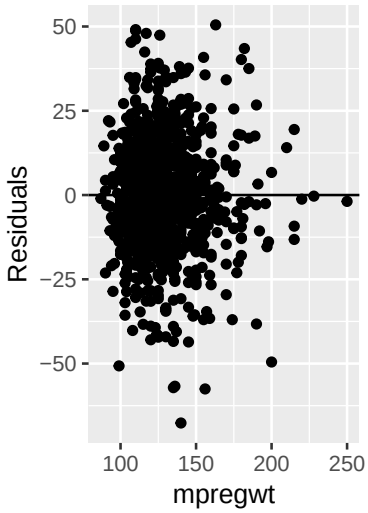


Fig 3e

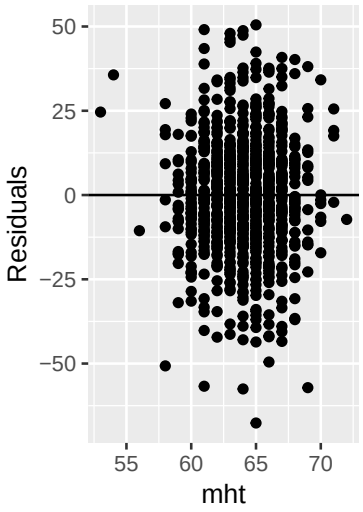


Fig 3f

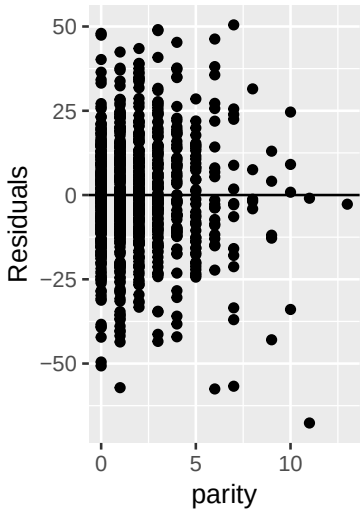


Fig 3g

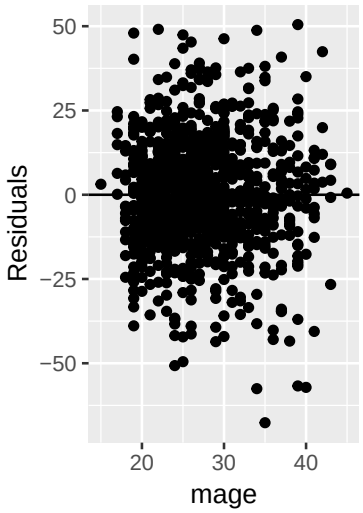
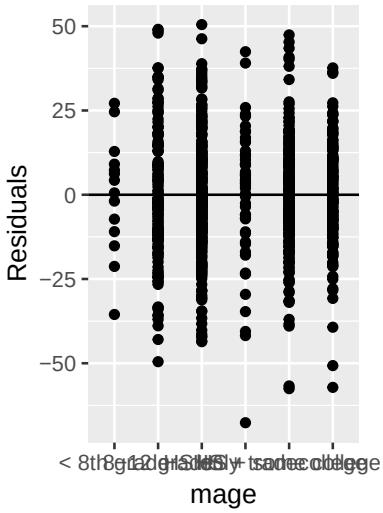


Fig 3h



Influential Points

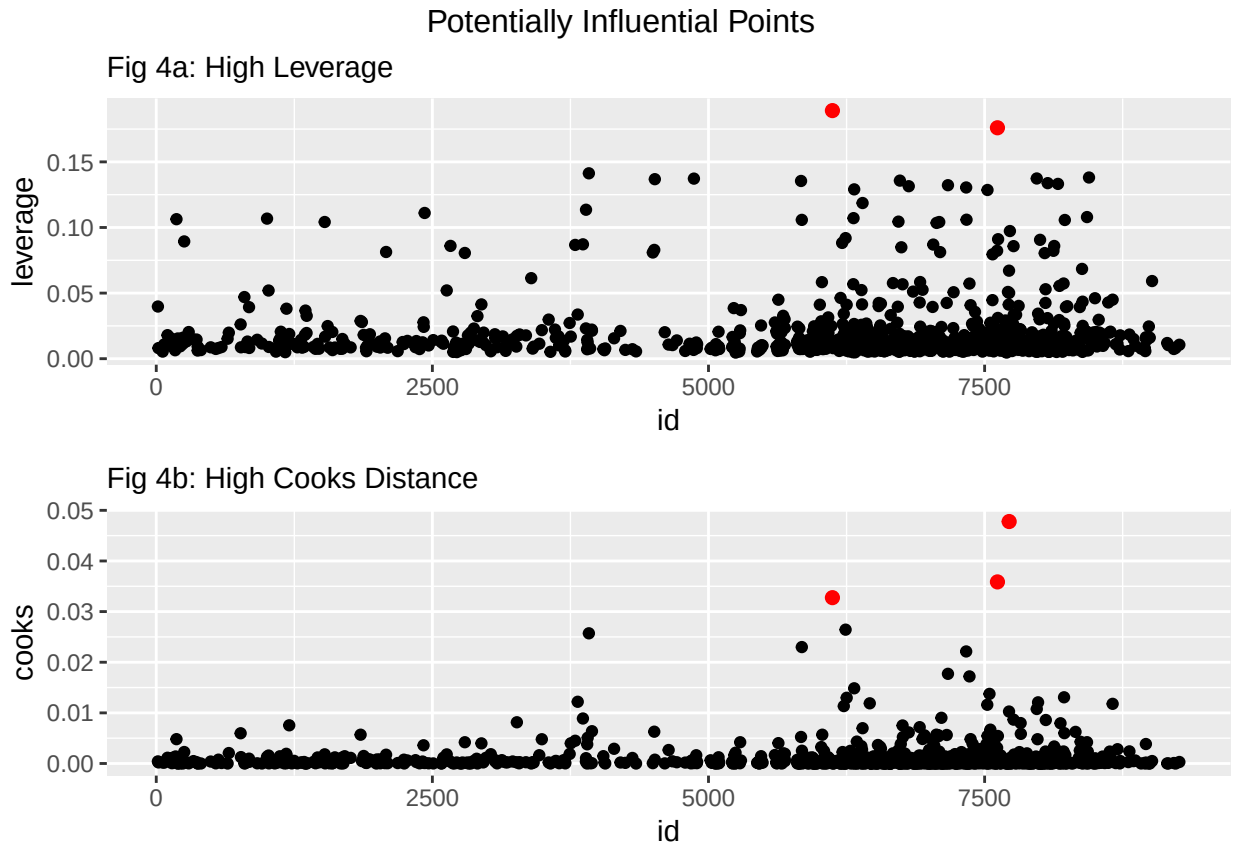
Before we finalize our model, we look for potentially influential points.

```
library(MASS)
```

```
Attaching package: 'MASS'
```

```
The following object is masked from 'package:dplyr':

    select
```

```
# Calcate 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)

# Plot leverage vs. id
l <- ggplot(data = smoke_leverage) +
    geom_point(mapping = aes(x = id, y = leverage)) +
    geom_point(data = smoke_leverage[smoke_leverage$leverage>.15, ],
               aes(x = id, y = leverage), color="red", size=2) +
    labs(subtitle = 'Fig 4a: High Leverage')
# Plot cooks vs. id
c <- ggplot(data = smoke_leverage) +
    geom_point(mapping = aes(x = id, y = cooks)) +
    geom_point(data = smoke_leverage[smoke_leverage$cooks>.03, ],
               aes(x = id, y = cooks), color="red", size=2) +
    labs(subtitle = 'Fig 4b: High Cooks Distance')

grid.arrange(l, c, top = 'Potentially Influential Points')
```




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

```
    id date bwt.oz parity mrace mage med mht mpregwt drace dage ded
1 7722 1563     55     11     7   35   3  65     140     7   36   4
2 6122 1713    146     10     6   39   0  53     110     6   41   0
3 7616 1638    144      0     6   27   0  58     102     6   27   1
  marital inc smoke  mraceF mpregwtC        mhtC      mageC        medF
1       1   6     0   black  11.4117   0.9537988  7.8398357  HS + trade
2       1   3     1 mexican -18.5883 -11.0462012 11.8398357 < 8th grade
3       1  NA     1 mexican -26.5883  -6.0462012 -0.1601643 < 8th grade
        medF2   leverage      cooks
1  HS + trade 0.05012875 0.04778706
2 < 8th grade 0.18908565 0.03275410
3 < 8th grade 0.17599724 0.03586226
```

We can see that the points with the highest leverage or cooks distance in our model are observations from mother's with babies who are either very light or very heavy at birth and that two of the three observations are from women with very high parity. Overall there observations are corner cases of our dataset and we remain confident that these points do not have a meaningful effect on our model. Therefore we select this model as our final model.

Discussion

Interpretation

Our final model has residual standard error of 16.732 and an R-squred of 0.176. The following table includes each elements point estimate, standard error, and 95% confidence interval.

```
summary(smoke_fit_final)
```

```

Call:
lm(formula = bwt.oz ~ smoke * mraceF + mraceF + mageC + mpregwtC +
    mhtC + parity + medF2, data = smoke)

Residuals:
   Min     1Q Median     3Q    Max
-67.64 -10.01  -0.56  10.29  50.48

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            124.40243    1.30941  95.006  < 2e-16 ***
smoke1                  -9.50320    1.30123  -7.303 5.92e-13 ***
mraceFblack            -11.43206    1.92027  -5.953 3.69e-09 ***
mraceFmexican            0.14556    3.84480   0.038 0.969809
mraceFasian             -5.57582    3.45002  -1.616 0.106388
mraceFmix               -0.15606    4.75307  -0.033 0.973814
mageC                   -0.09102    0.11912  -0.764 0.444988
mpregwtC                 0.11568    0.02998   3.859 0.000122 ***
mhtC                     0.98116    0.25207   3.892 0.000106 ***
parity                   0.95854    0.36017   2.661 0.007913 **
medF2< 8th grade        -7.72820    4.78226  -1.616 0.106421
medF28-12 grade         -3.73728    1.72824  -2.162 0.030829 *
medF2HS + trade         -2.41855    2.53219  -0.955 0.339758
medF2HS + some college   0.82463    1.44255   0.572 0.567695
medF2college            -1.64162    1.64855  -0.996 0.319601
smoke1:mraceFblack       2.98799    2.75186   1.086 0.277838
smoke1:mraceFmexican    18.54938    7.15273   2.593 0.009651 **
smoke1:mraceFasian      -9.12004    6.33960  -1.439 0.150597
smoke1:mraceFmix        -8.84448    7.65275  -1.156 0.248083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.73 on 955 degrees of freedom
Multiple R-squared:  0.1763,    Adjusted R-squared:  0.1607
F-statistic: 11.35 on 18 and 955 DF,  p-value: < 2.2e-16
```

```
coeff <- cbind(summary(smoke_fit_final)$coefficients[, c(1,2)],
               confint(smoke_fit_final))
coeff <- data.frame(coeff)
names(coeff) <- c('Estimate', 'Std.Error', '2.5%', '97.5%')
coeff
```

```
                          Estimate  Std.Error         2.5%       97.5%
(Intercept)            124.40243028 1.30941113 121.83277493 126.9720856
smoke1                  -9.50319924 1.30122864 -12.05679685  -6.9496016
```

```
mraceFblack            -11.43205650 1.92026517 -15.20048307  -7.6636299
mraceFmexican            0.14555791 3.84480104  -7.39967626   7.6907921
mraceFasian             -5.57581992 3.45002080 -12.34631716   1.1946773
mraceFmix               -0.15606193 4.75306505  -9.48371983   9.1715960
mageC                   -0.09101835 0.11911625  -0.32477818   0.1427415
mpregwtC                 0.11568152 0.02997949   0.05684823   0.1745148
mhtC                     0.98115803 0.25206677   0.48648931   1.4758268
parity                   0.95854043 0.36016695   0.25173038   1.6653505
medF2< 8th grade        -7.72819667 4.78225599 -17.11314037   1.6567470
medF28-12 grade         -3.73727695 1.72824296  -7.12886931  -0.3456846
medF2HS + trade         -2.41854826 2.53218802  -7.38784351   2.5507470
medF2HS + some college   0.82462903 1.44254819  -2.00630130   3.6555594
medF2college            -1.64162421 1.64855168  -4.87682633   1.5935779
smoke1:mraceFblack       2.98799067 2.75186080  -2.41240167   8.3883830
smoke1:mraceFmexican    18.54937905 7.15273380   4.51248849  32.5862696
smoke1:mraceFasian      -9.12003652 6.33960459 -21.56120074   3.3211277
smoke1:mraceFmix        -8.84448131 7.65275302 -23.86263517   6.1736726
```

Intercept: Babies for a white, non-smoking, women of average height and weight, with no previous pregnancies with a high school only education have an estimated average birth weight of 124.4 ozs (95% CI: 121.8, 127)

Smoking: Holding all else constant, women who have smoked at some point in her life are estimated to have an average baby's birth weight to decrease -9.5 ozs (95% CI: -12.06, -6.95).

Mother's Pre-pregnancy Weight: Holding all else constant, for each additional lb mothers weighed before pregnancy, we estimate average baby's birth weight to increase 0.12 ozs (95% CI: 0.06, 0.17).

Mother's Height: Holding all else constant, for each additional inch mothers have in height, we expect average baby's birth weight to increase 0.98 ozs (95% CI: 0.49, 1.48).

Mother's Age: Holding all else constant, for each additional year mothers age we expect average baby's birth weight to decrease -0.09 ozs (95% CI: -0.32, 0.14). Given that this confidence interval includes 0, we are not confident that there is a meaningful effect of mother's age on birth weight.

Parity: Holding all else constant, for each additional pregnancy mothers had before the current pregnancy we expect average baby's birth weight to increase 0.96 ozs (95% CI: 0.25, 1.67).

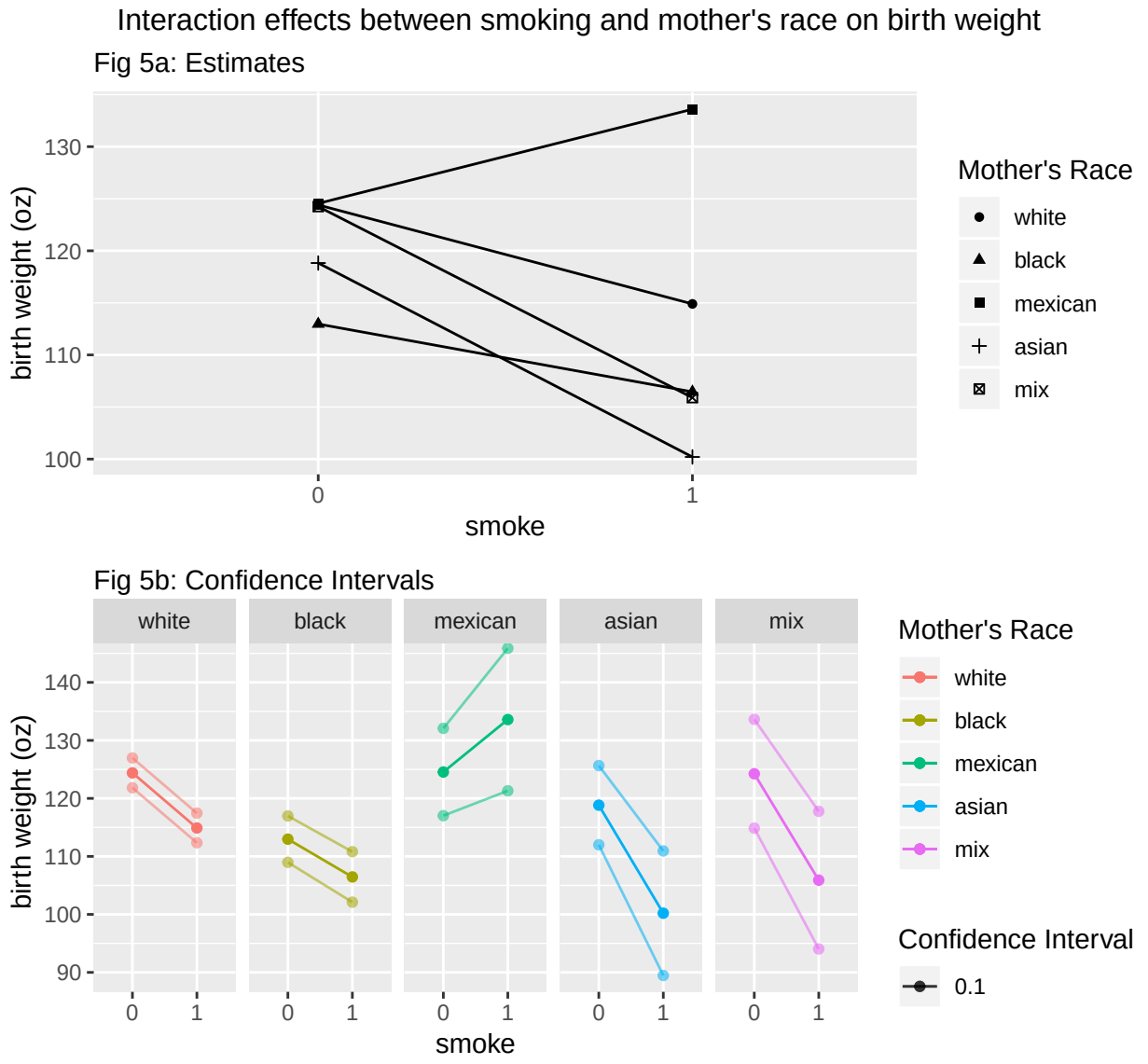
Education Holding all else constant, for a woman with:

- Less than an 8th grade education we estimate average birth weights to be -7.73 ozs less than a woman with a high school only education, (95% CI:-17.11, -0.35)
- Between an 8th grade and a 12th grade education we estimate average birth weights to be -3.74 ozs less than a woman with a high school only education, (95% CI:-7.13, -0.35)
- High school and trade school education we estimate average birth weights to be -2.42 ozs less than a woman with a high school only education, (95% CI:-7.39, 2.55). Given that this confidence interval includes 0, we are not confident that there is a meaningful difference.
- High school and some college education we estimate average birth weights to be 0.82 ozs more than a woman with a high school only education, (95% CI:-2.01, 3.66). Given that this confidence interval includes 0, we are not confident that there is a meaningful difference.
- College education we estimate average birth weights to be -1.64 ozs less than a woman with a high school only education, (95% CI:-4.88, 1.59). Given that this confidence interval includes 0, we are not confident that there is a meaningful difference.

Mother's Race Holding all else constant, we see race effecting birth weight in the following way (Figure 5):

```r
# Rebase our models so we can discuss the impacts of smoking
# for each race individually
# Black
smoke_black <- smoke %>%
    mutate(mraceF = factor(mraceF, levels = c('black', 'white', 'mexican',
                                              'asian', 'mix')))
smoke_fit_black <- lm(bwt.oz ~ smoke*mraceF + mraceF + mageC + mpregwtC +
                          mhtC + parity + medF2, data = smoke_black)
coeff_black <- cbind(summary(smoke_fit_black)$coefficients[, c(1,2)],
               confint(smoke_fit_black))
coeff_black <- data.frame(coeff_black)
names(coeff_black) <- c('Estimate', 'Std.Error', '2.5%', '97.5%')

# Mexican
smoke_mex <- smoke %>%
    mutate(mraceF = factor(mraceF, levels = c('mexican', 'white', 'black',
                                              'asian', 'mix')))
smoke_fit_mex <- lm(bwt.oz ~ smoke*mraceF + mraceF + mageC + mpregwtC +
                          mhtC + parity + medF2, data = smoke_mex)
coeff_mex <- cbind(summary(smoke_fit_mex)$coefficients[, c(1,2)],
               confint(smoke_fit_mex))
coeff_mex <- data.frame(coeff_mex)
names(coeff_mex) <- c('Estimate', 'Std.Error', '2.5%', '97.5%')

# Asian
smoke_asia <- smoke %>%
    mutate(mraceF = factor(mraceF, levels = c('asian', 'white', 'mexican',
                                              'black', 'mix')))
smoke_fit_asia <- lm(bwt.oz ~ smoke*mraceF + mraceF + mageC + mpregwtC +
                          mhtC + parity + medF2, data = smoke_asia)
coeff_asia <- cbind(summary(smoke_fit_asia)$coefficients[, c(1,2)],
               confint(smoke_fit_asia))
coeff_asia <- data.frame(coeff_asia)
names(coeff_asia) <- c('Estimate', 'Std.Error', '2.5%', '97.5%')

# Mix
smoke_mix <- smoke %>%
    mutate(mraceF = factor(mraceF, levels = c('mix', 'white', 'mexican',
                                              'asian', 'black')))
smoke_fit_mix <- lm(bwt.oz ~ smoke*mraceF + mraceF + mageC + mpregwtC +
                          mhtC + parity + medF2, data = smoke_mix)
coeff_mix <- cbind(summary(smoke_fit_mix)$coefficients[, c(1,2)],
               confint(smoke_fit_mix))
coeff_mix <- data.frame(coeff_mix)
names(coeff_mix) <- c('Estimate', 'Std.Error', '2.5%', '97.5%')

# Create new temporary dataset for confidence interval
newvals <- data.frame(smoke = rep(c(0,1),5),
                      mraceF = c(rep('white', 2),
                                 rep('black', 2),
                                 rep('mexican', 2),
                                 rep('asian', 2),
                                 rep('mix', 2)),
```

```
                      mageC = 0,
                      mpregwtC = 0,
                      mhtC = 0,
                      parity = 0,
                      medF2 = 'HS only') %>%
    mutate(smoke = as.factor(smoke),
           mraceF = factor(mraceF, levels = c('white', 'black', 'mexican',
                                              'asian', 'mix')))
# Create confidence interval
predict <- predict.lm(smoke_fit_final, newvals, interval = 'confidence')
# Append confidence interval to temporary dataset
newvals <- newvals %>%
    mutate(fit = predict[,1],
           lwr = predict[,2],
           upr = predict[,3])

# Plot the interaction effects between smoking and mother's race on birth weight
p1 <- ggplot(data = newvals) +
    geom_point(mapping = aes(x = smoke, y = fit, group = mraceF,
                                    shape = mraceF)) +
    geom_line(mapping = aes(x = smoke, y = fit, group = mraceF)) +
    labs(subtitle = 'Fig 5a: Estimates', shape = 'Mother\'s Race') +
    ylab('birth weight (oz)')


# Plot the confidence interval
p2 <- ggplot(data = newvals) +
    geom_point(mapping = aes(x = smoke, y = fit, group = mraceF,
                                    color = mraceF)) +
    geom_line(mapping = aes(x = smoke, y = fit, group = mraceF,
                                    color = mraceF)) +
    geom_point(mapping = aes(x = smoke, y = lwr, group = mraceF,
                                    color = mraceF, alpha = .1)) +
    geom_line(mapping = aes(x = smoke, y = lwr, group = mraceF,
                                    color = mraceF, alpha = .1)) +
    geom_point(mapping = aes(x = smoke, y = upr, group = mraceF,
                                    color = mraceF, alpha = .1)) +
    geom_line(mapping = aes(x = smoke, y = upr, group = mraceF,
                                    color = mraceF, alpha = .1)) +
    facet_grid(. ~ mraceF) +
    labs(subtitle = 'Fig 5b: Confidence Intervals', color = 'Mother\'s Race',
         alpha = 'Confidence Interval') +
    ylab('birth weight (oz)')

grid.arrange(p1, p2,
             top = 'Interaction effects between smoking and mother\'s race on birth weight')
```



Conclusion

Thinking back to our original research questions. In general, our findings suggest, for most demographics, mothers who smoke do tend to give birth to babies with lower weights than mothers who do not smoke. However, our findings also suggest that this relationship differs by mothers race. Specifically,

- White women who smoke are estimated to have babies, on average, -9.5 ozs lighter (95% CI: -12.06, -6.95) than White women who don't smoke.
- Black women who smoke are estimated to have babies, on average, -6.52 ozs lighter (95% CI: -11.3, -1.73) than Black women who don't smoke.
- Asian women who smoke are estimated to have babies, on average, -18.62 ozs lighter (95% CI: -30.81, -6.44) than Asian women who don't smoke.
- Multiracial women who smoke are estimated to have babies, on average, -18.35 ozs lighter (95% CI: -33.16, -3.54) than multical women who don't smoke.
- Mexican women who smoke are estimated to have babies, on average, 9.05 ozs heavier (95% CI: -4.76, 22.85) than 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 is Mexican women, although our estimates suggest that smoking might actually have a positive relationship to birth weight for Mexican women. However, it should be noted that Mexican women only accounted for 3% of our dataset (29 observations). Furthermore, our confidence intervals for this coefficient included zero, 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.

Beyond the unique relationship between smoking and birth weight in Mexican women, our findings also suggest that smoking may have heavier effects on birth weight in Asian and Multiracial woman compared to White and Black women. Again, it should be noted that Asian and Multiracial women only accounted for 3.8% and 2.2% of our dataset respectively. Therefore additional research is needed to understand the strength of the relationship between smoking and birth weight in a diverse population of women.

Another interesting finding has to do with parity. Our results suggest that increased total number of previous pregnancies, including fetal deaths and still births is associated with increased birth weights. Specifically, for each additional pregnancy a mother had before the current pregnancy we expect average birth weights to increase 0.96 ozs (95% CI: 0.25, 1.67). Intuitively, this researcher would not expect previous pregnancies to have an effect on future pregnancy baby weight. However, our research suggests that that average weights of babies should increase with the number of previous pregnancies. Future research should examine the reasons behind this association.

Yet another interesting finding is around mother's education. Our results suggest that there may be slight differences in birth weight associated with differing levels of mother's education. Specifically, it seems an education that did not include the completion of high school is associated with lower birth weights that those who did complete high school and/or additional schooling. One explanation for this finding could be tied to socioeconomic status and other environmental factors beyond education. Mothers who did not complete high school might find themselves with fewer resources to sustain optimal health for both themselves and their children than do mothers with higher levels of education. It should be noted that our results are, in fact, very small with many factors showing insignificant differences. Nevertheless, further research should be done on the mediating factors between mother's education and birth weight.

Limitations

Our final model has residual standard error of 16.732 and an R-squred of 0.2 meaning that our model accounts for only 17.626% of the variation of birth weight. It seems that we are missing variables in our model that would explain additional variation in birth weight, 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. 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.