Homework3

Echo Liu

September 25, 2018

Homework 3: Maternal Smoking and Birth Weights

Our goal is to study whether maternal smoking would affect birth weights given outcome variable body weights and predictors like mother's weight and height, mother's race, mother's smoking habit and even father's information. I decided to use the cleaned data which excludes all missing values and all of the variables on the fathers because of the missing information situation in fathers' weights and heights and potential multicolinearity problem. Let's first read in the data and look at summary of this data set.

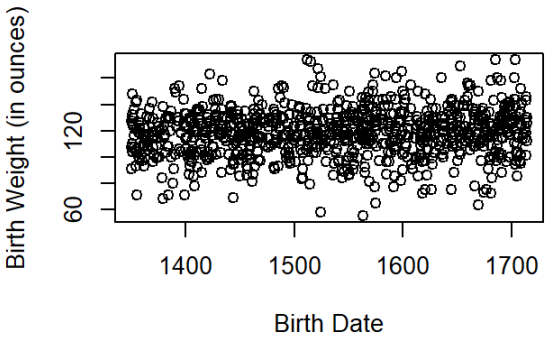
```
##        id            date        gestation         bwt.oz
##  Min.   :  15   Min.   :1350   Min.   :148.0   Min.   : 55.0
##  1st Qu.:5477   1st Qu.:1444   1st Qu.:272.0   1st Qu.:108.0
##  Median :6734   Median :1540   Median :279.0   Median :119.0
##  Mean   :6032   Mean   :1536   Mean   :278.5   Mean   :118.4
##  3rd Qu.:7587   3rd Qu.:1627   3rd Qu.:286.0   3rd Qu.:129.0
##  Max.   :9263   Max.   :1714   Max.   :338.0   Max.   :174.0
##      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   : 1.953   Mean   :2.995   Mean   :27.29   Mean   :2.932
##  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.   :7.000
##       mht           mpregwt           inc            smoke
##  Min.   :53.00   Min.   : 87.0   Min.   :0.000   Min.   :0.0000
##  1st Qu.:62.00   1st Qu.:113.0   1st Qu.:2.000   1st Qu.:0.0000
##  Median :64.00   Median :125.0   Median :3.000   Median :0.0000
##  Mean   :64.07   Mean   :128.5   Mean   :3.681   Mean   :0.4638
##  3rd Qu.:66.00   3rd Qu.:140.0   3rd Qu.:5.000   3rd Qu.:1.0000
##  Max.   :72.00   Max.   :220.0   Max.   :9.000   Max.   :1.0000
```

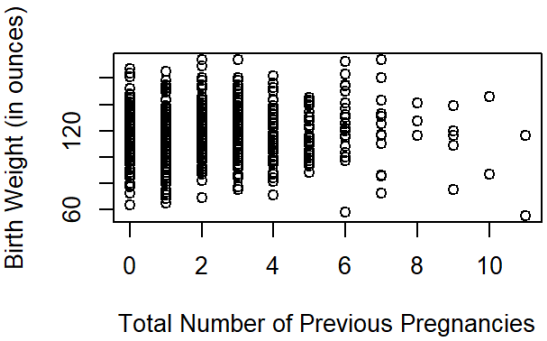
I decided to drop unrelated columns: id and gestation from the original table. Also I renamed race category from number to actual race and grouped mrace from 0-5 into "white".

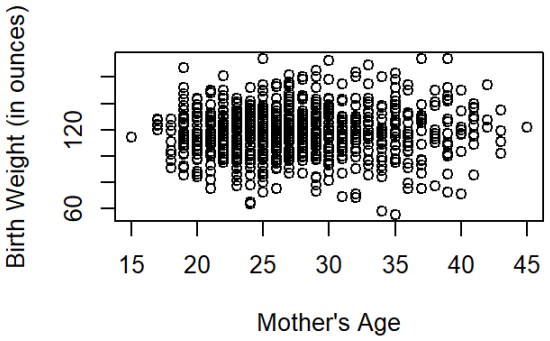
```
#drop unnecessary columns
babies$id = NULL
babies$gestation = NULL

#deal with 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"

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

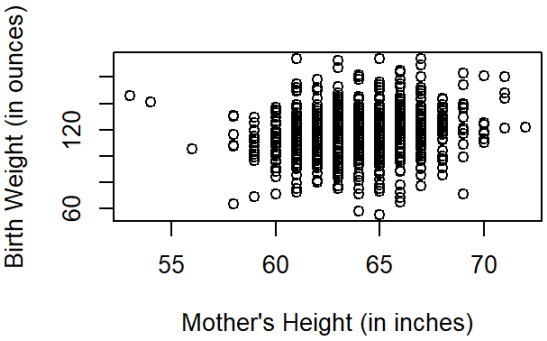
Exploratory Data Analysis

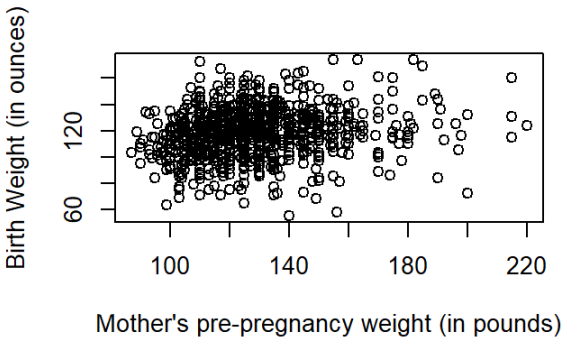
Scatter plots and Box Plots

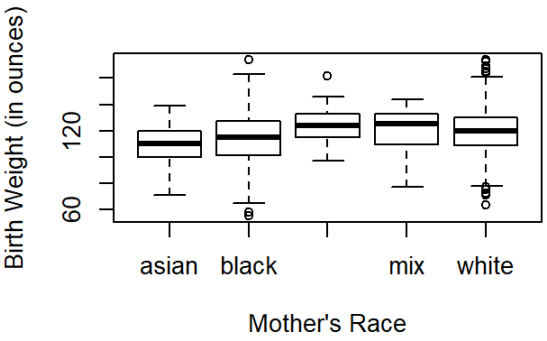

Birth Weight (in ounces)
60
120
1400
1500
1600
1700
Birth Date

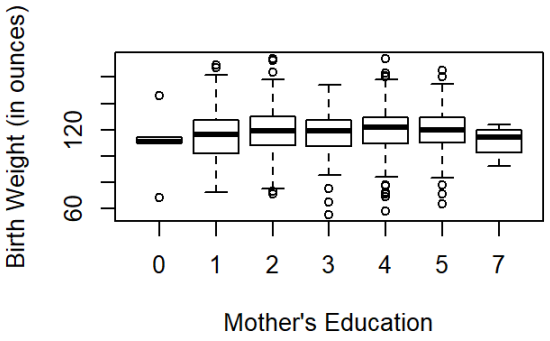

Birth Weight (in ounces)
60
120
0
2
4
6
8
10
Total Number of Previous Pregnancies

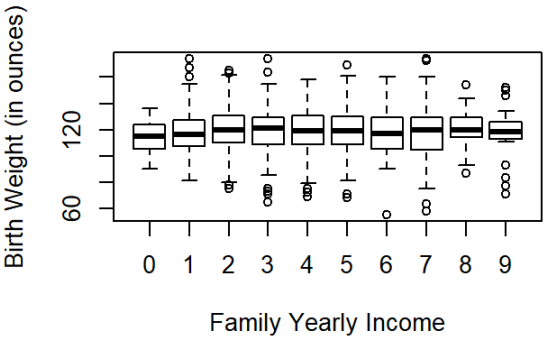

Birth Weight (in ounces)
60
120
15
20
25
30
35
40
45
Mother's Age

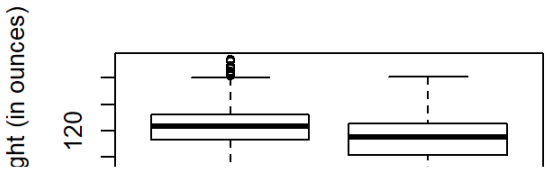

Birth Weight (in ounces)
60
120
55
60
65
70
Mother's Height (in inches)

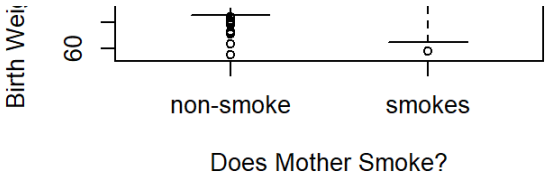

Birth Weight (in ounces)
60
120
100
140
180
220
Mother's pre-pregnancy weight (in pounds)


Birth Weight (in ounces)
60
120
asian
black
mix
white
Mother's Race


Birth Weight (in ounces)
60
120
0
1
2
3
4
5
7
Mother's Education


Birth Weight (in ounces)
60
120
0
1
2
3
4
5
6
7
8
9
Family Yearly Income


ght (in ounces)
120



No real patterns show up in the scatter plots and constant variance assumptions seem to be satisfied in all plots except for boxplot for babies' weights versus mother's education. Variance box around "0" category is less than half of category "1". Let's look into the "0" category of mother's education:

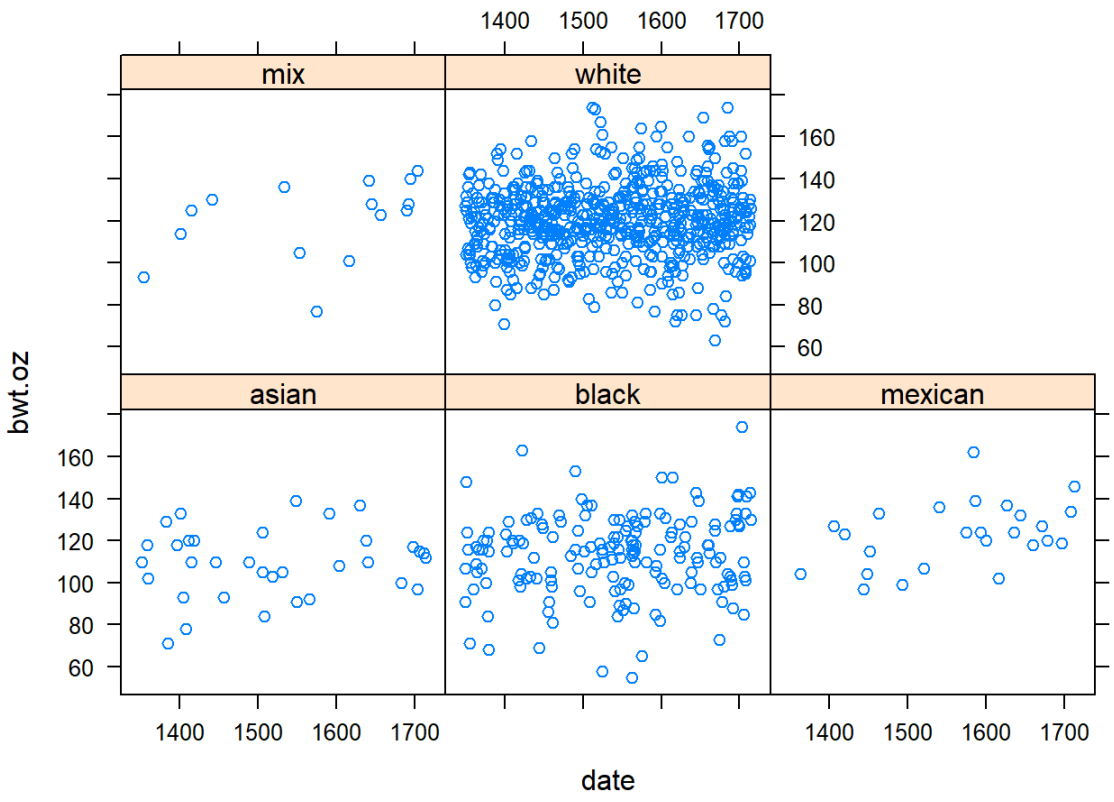
```
table(babies$med)
```

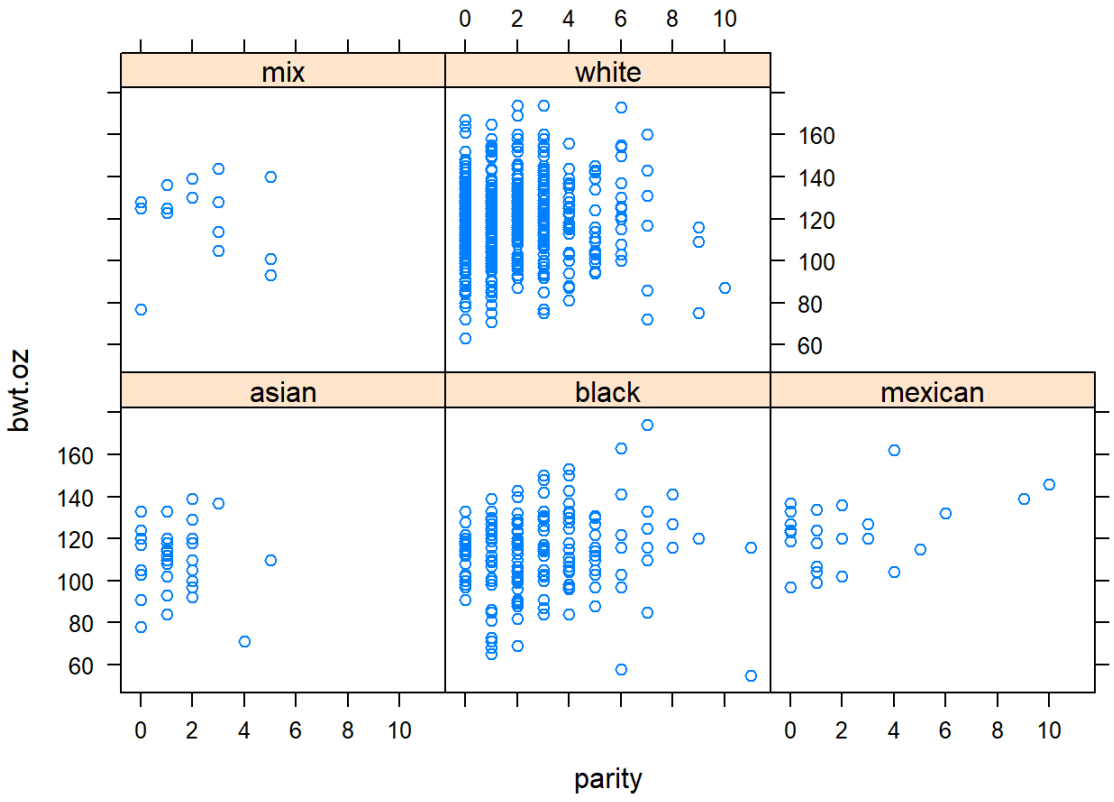
```
##
##   0   1   2   3   4   5   7
##   5 130 321  47 203 159   4
```

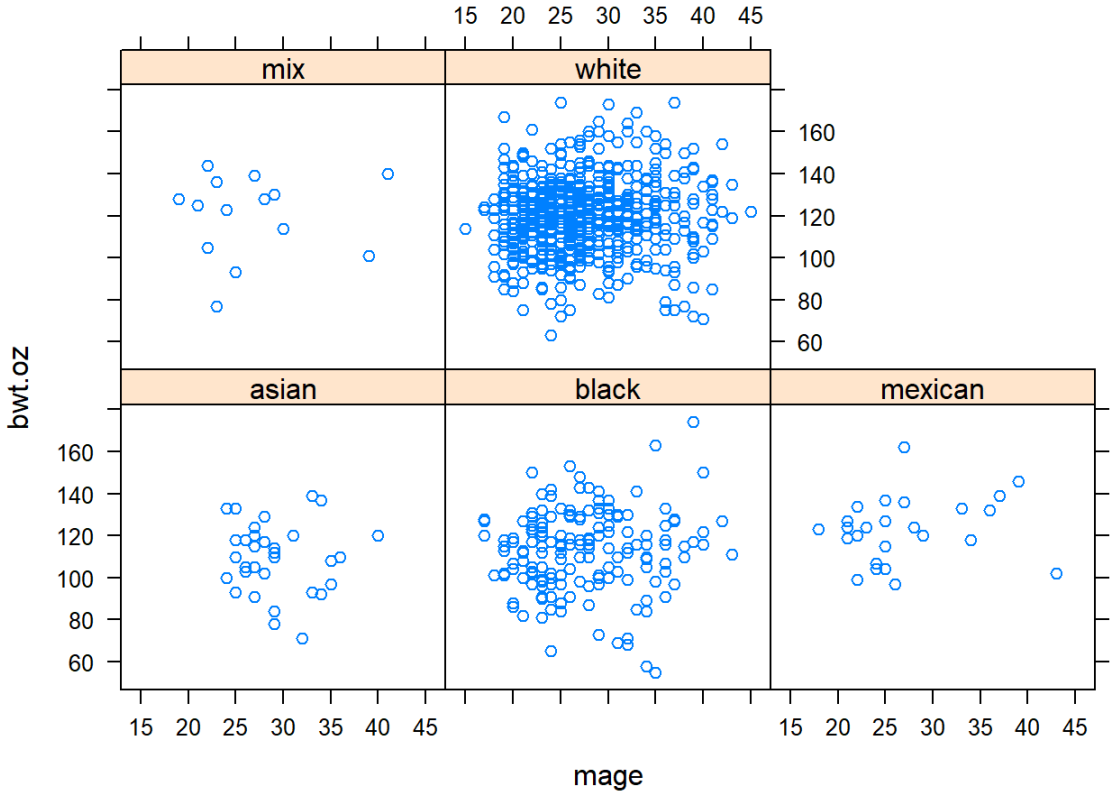
The reason why there seems to be a non-constant problem is because of lack of data (or there are few instances) in the category where mother's education equals 0.

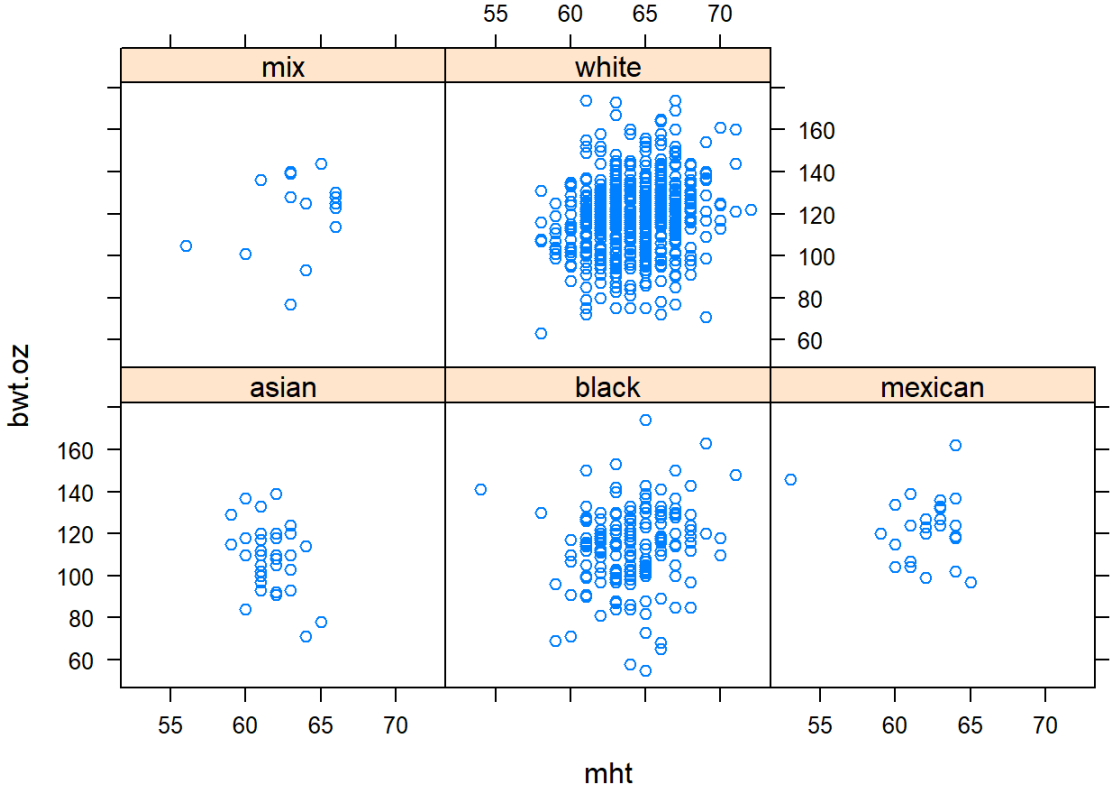
Interactions

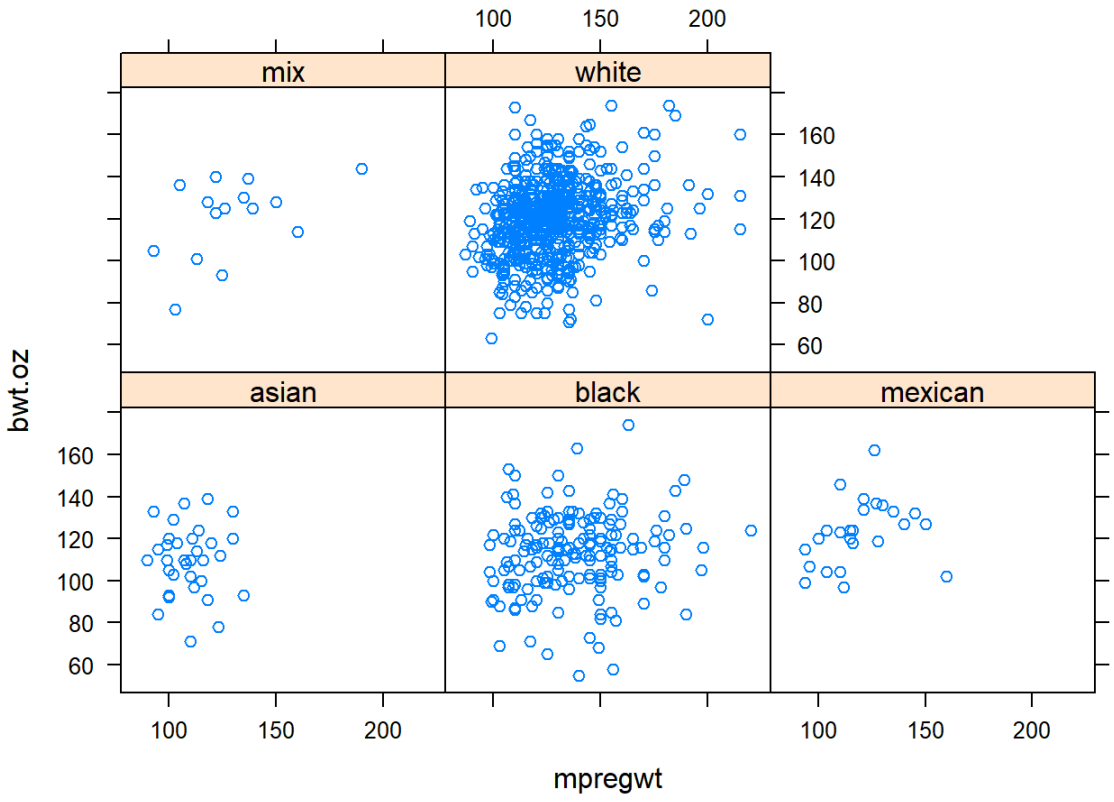
Let’s see if there are any interactions with race, which is an interesting variable.

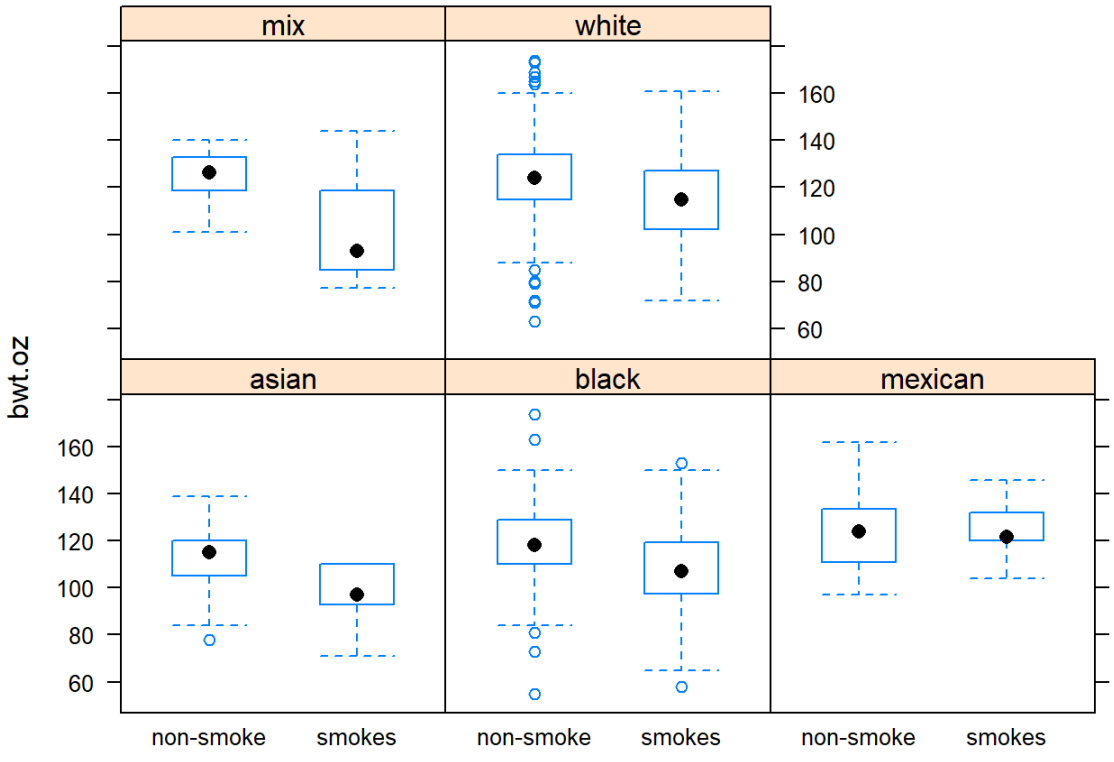


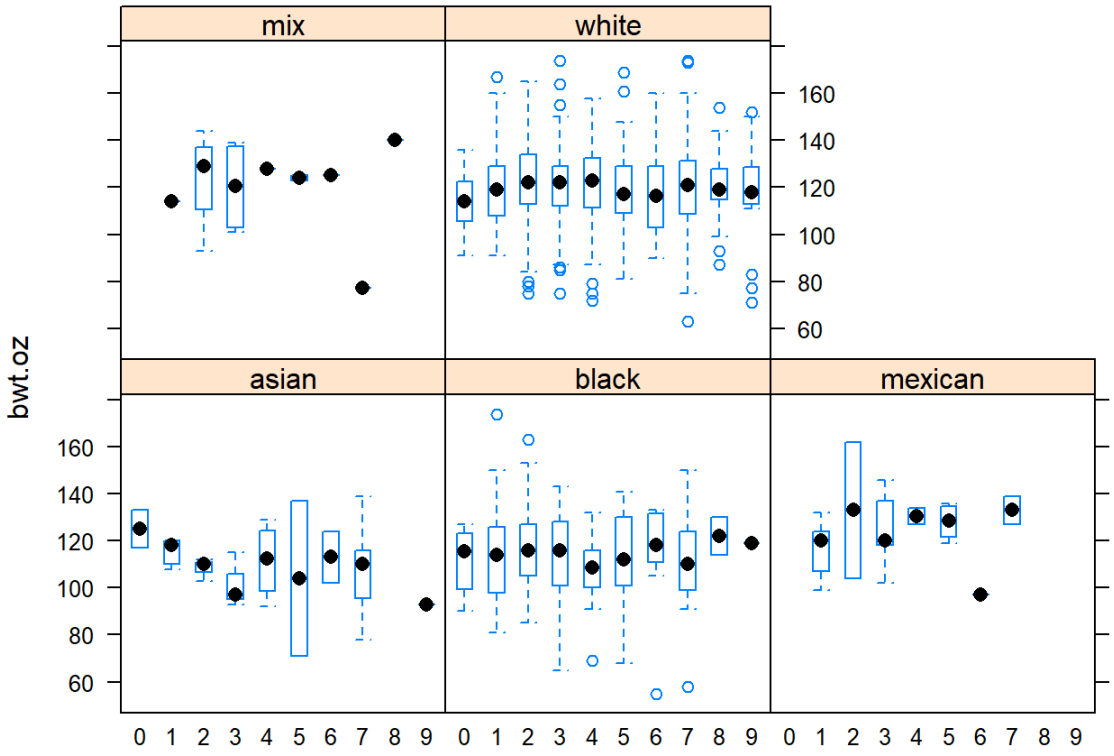


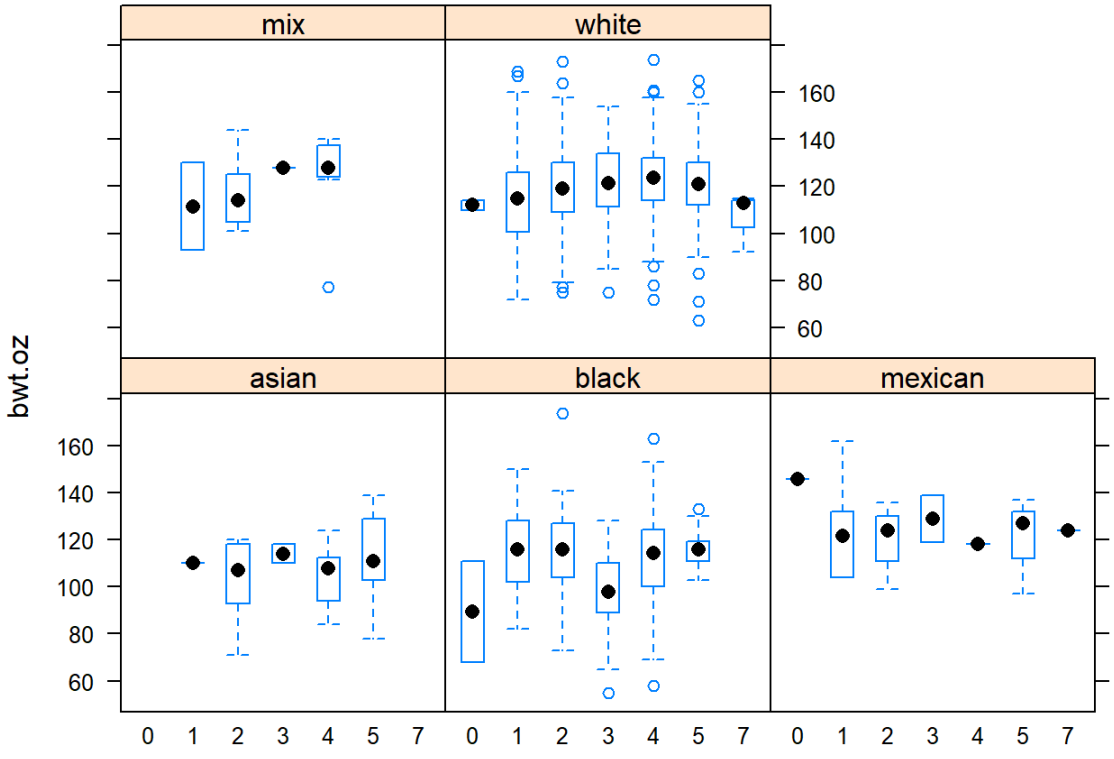

mix
white
asian
black
mexican
bwt.oz
mage
15 20 25 30 35 40 45
60 80 100 120 140 160


mix
white
asian
black
mexican
bwt.oz
mht
55 60 65 70
60 80 100 120 140 160


mix
white
asian
black
mexican
bwt.oz
mpregwt
100
150
200
60
80
100
120
140
160


mix
white
asian
black
mexican
bwt.oz
non-smoke
smokes
60
80
100
120
140
160





The only interaction we can detect from the plot is interaction between race and mother's height since the slope in "asian" subplot seems to negative while slopes in other subplots tend to be positive.

Let's also explore whether there exists other interactions. Unfortuately, there is no obvious interaction in any other plots.

```
xyplot(bwt.oz ~ date | med, data = babies)
xyplot(bwt.oz ~ parity | med, data = babies)
xyplot(bwt.oz ~ mage | med, data = babies)
xyplot(bwt.oz ~ mht | med, data = babies)
xyplot(bwt.oz ~ mpregwt | med, data = babies)
bwplot(bwt.oz ~ smokef | as.factor(med), data = babies)
bwplot(bwt.oz ~ as.factor(inc) | as.factor(med), data = babies)


xyplot(bwt.oz ~ date | smokef, data = babies)
xyplot(bwt.oz ~ parity | smokef, data = babies)
xyplot(bwt.oz ~ mage | smokef, data = babies)
xyplot(bwt.oz ~ mht | smokef, data = babies)
xyplot(bwt.oz ~ mpregwt | smokef, data = babies)
bwplot(bwt.oz ~ as.factor(inc) | smokef, data = babies)

xyplot(bwt.oz ~ date | as.factor(inc), data = babies)
xyplot(bwt.oz ~ parity | as.factor(inc), data = babies)
xyplot(bwt.oz ~ mage | as.factor(inc), data = babies)
xyplot(bwt.oz ~ mht | as.factor(inc), data = babies)
xyplot(bwt.oz ~ mpregwt | as.factor(inc), data = babies)

bwplot(bwt.oz ~ as.factor(inc) | as.factor(med), data = babies)
```

Check if there is multicollinearity problem

Take a look at correlations among predictors for multicollinearity.

```
##           date parity  mrace   mage    med    mht mpregwt    inc  smoke
## date     1.000  0.080  0.056  0.065 -0.066 -0.066   0.025  0.067 -0.079
## parity   0.080  1.000  0.149  0.524 -0.201 -0.043   0.151  0.009  0.011
## mrace    0.056  0.149  1.000  0.014 -0.079 -0.165   0.023 -0.122 -0.114
## mage     0.065  0.524  0.014  1.000  0.134 -0.005   0.146  0.297 -0.070
## med     -0.066 -0.201 -0.079  0.134  1.000  0.115  -0.054  0.217 -0.138
## mht     -0.066 -0.043 -0.165 -0.005  0.115  1.000   0.460  0.071  0.041
## mpregwt  0.025  0.151  0.023  0.146 -0.054  0.460   1.000 -0.005 -0.049
## inc      0.067  0.009 -0.122  0.297  0.217  0.071  -0.005  1.000  0.007
## smoke   -0.079  0.011 -0.114 -0.070 -0.138  0.041  -0.049  0.007  1.000
```

There is few high correlation among the predictors. The only two relatively significant correlations are 0.524, which is between mage and parity, and 0.460 which is between mht and mpregwt. Multicollinearity problem doesn't necessarily exist because correlations are not as high as 0.9. Thus we still keep all predictors in the model.

Modelling

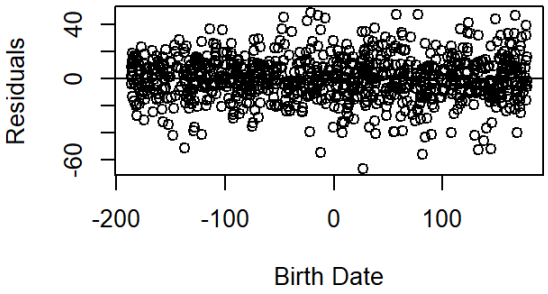
```
#let's center all the continuous predictors to get a better interpretation for the intercept
babies$cdate = babies$date - mean(babies$date)
babies$cparity = babies$parity - mean(babies$parity)
babies$cmage = babies$mage - mean(babies$mage)
babies$cmht = babies$mht - mean(babies$mht)
babies$cmpregwt = babies$mpregwt - mean(babies$mpregwt)

#do a quick check to make sure the code did what we wanted it to do
head(babies)
```

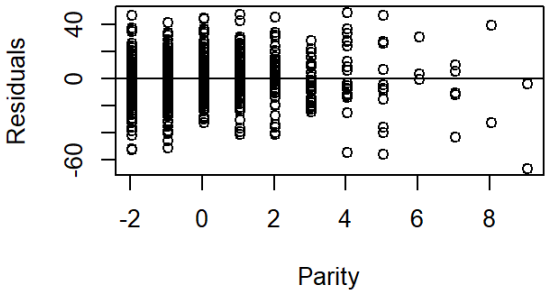
```
##   date bwt.oz parity mrace mage med mht mpregwt inc smoke mracef    smokef
## 1 1598    116      7     7   28   1  66     135   2     0  black non-smoke
## 2 1527    110      7     7   27   1  64     133   1     0  black non-smoke
## 3 1563     55     11     7   35   3  65     140   6     0  black non-smoke
## 4 1503    132      4     7   28   2  67     148   3     0  black non-smoke
## 5 1638    105      4     7   34   3  61     130   3     0  black non-smoke
## 6 1705     85      7     7   33   1  67     130   2     0  black non-smoke
##        cdate  cparity      cmage        cmht  cmpregwt
## 1  61.576525 5.047181  0.7054085  1.93095512  6.521289
## 2  -9.423475 5.047181 -0.2945915 -0.06904488  4.521289
## 3  26.576525 9.047181  7.7054085  0.93095512 11.521289
## 4 -33.423475 2.047181  0.7054085  2.93095512 19.521289
## 5 101.576525 2.047181  6.7054085 -3.06904488  1.521289
## 6 168.576525 5.047181  5.7054085  2.93095512  1.521289
```

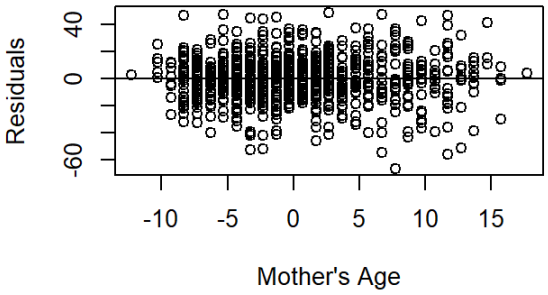
First model: plain vanilla

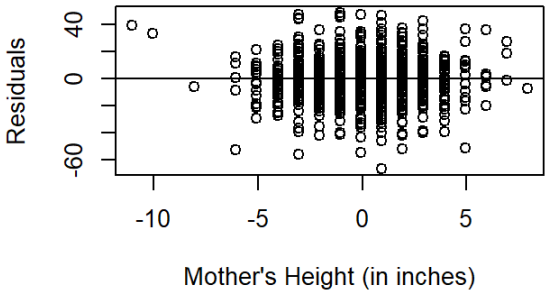
```
reg1 <- lm(bwt.oz ~ cdate + cparity + cmage + cmht + cmpregwt + as.factor(mracef) + as.factor(med) + as.fac
tor(inc) + as.factor(smokef), data = babies)
```

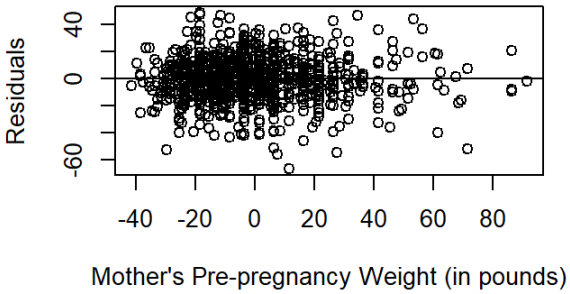
To do check of residuals versus each predictor to make sure validity of assumptions are satisfied:

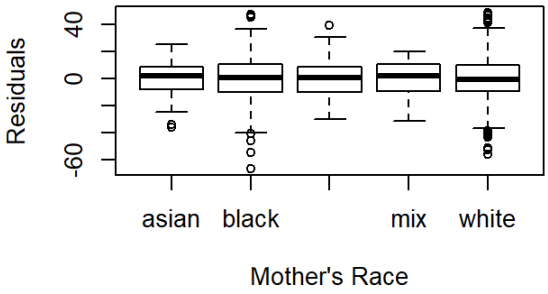

Residuals
40
0
-60
-200
-100
0
100
Birth Date

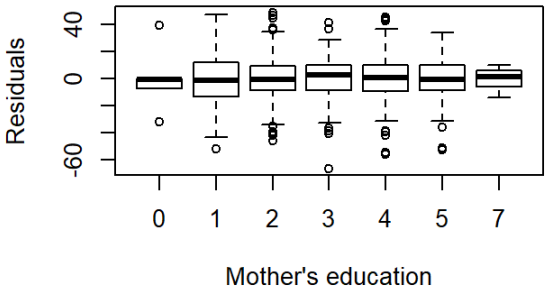

Residuals
40
0
-60
-2
0
2
4
6
8
Parity

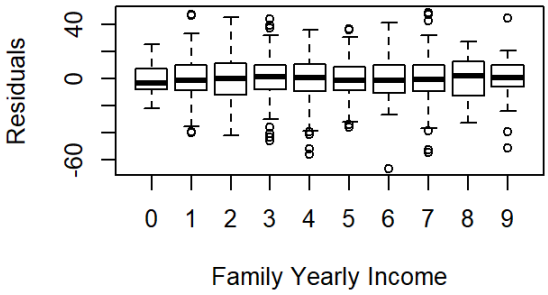

Residuals
40
0
-60
-10
-5
0
5
10
15
Mother's Age

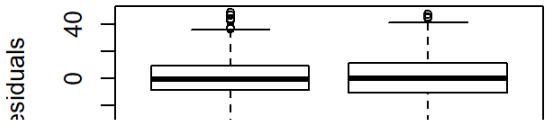

Residuals
40
0
-60
-10
-5
0
5
Mother's Height (in inches)

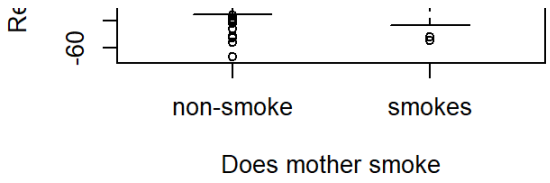

Residuals
40
0
-60
-40
-20
0
20
40
60
80
Mother's Pre-pregnancy Weight (in pounds)


Residuals
40
0
-60
asian
black
mix
white
Mother's Race


Residuals
40
0
-60
0
1
2
3
4
5
7
Mother's education


Residuals
40
0
-60
0
1
2
3
4
5
6
7
8
9
Family Yearly Income


esiduals
40
0



All plots look good (satisfied linearity and constant-variance assumption) except for mother's education (doesn't seem to satisfy constant-variance assumption). Thus we look into its summary table:

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

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

The reason why variance around "0" category seems to be half of variance around "1" category is because of the lack of data (or just few instances in reality) in female who has 0 years of schooling. Therefore, we can assume constant-variance assumption is satisfied in this case. And also I notice two outliers in residual plot for parity. We might look into this in the leverage point section.

Let's look at summary of this model.

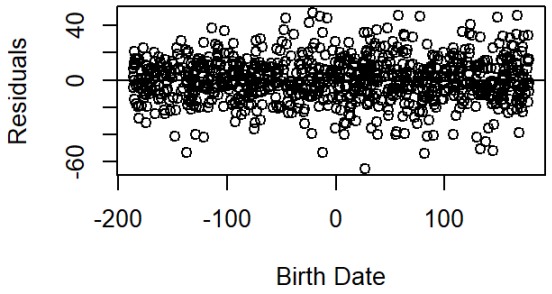
```
##
## Call:
## lm(formula = bwt.oz ~ cdate + cparity + cmage + cmht + cmpregwt +
##     as.factor(mracef) + as.factor(med) + as.factor(inc) + as.factor(smokef),
##     data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.639  -9.423  -0.108  10.078  48.796
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               107.878548   8.632618  12.497  < 2e-16 ***
## cdate                       0.013113   0.005454   2.404 0.016418 *
## cparity                     0.765674   0.399790   1.915 0.055806 .
## cmage                      -0.035073   0.133964  -0.262 0.793535
## cmht                        0.989393   0.269550   3.671 0.000257 ***
## cmpregwt                    0.106565   0.032886   3.240 0.001240 **
## as.factor(mracef)black     -1.787937   3.407244  -0.525 0.599898
## as.factor(mracef)mexican   11.330692   4.544646   2.493 0.012851 *
## as.factor(mracef)mix        4.404848   5.348201   0.824 0.410392
## as.factor(mracef)white      7.451374   3.115701   2.392 0.016995 *
## as.factor(med)1             5.208750   7.793545   0.668 0.504098
## as.factor(med)2             7.574706   7.687072   0.985 0.324719
## as.factor(med)3             5.960238   7.980229   0.747 0.455347
## as.factor(med)4             8.149708   7.731512   1.054 0.292144
## as.factor(med)5             7.190699   7.763601   0.926 0.354604
## as.factor(med)7            -4.190303  11.297513  -0.371 0.710802
## as.factor(inc)1             3.056688   3.585926   0.852 0.394227
## as.factor(inc)2             4.601360   3.603853   1.277 0.202028
## as.factor(inc)3             1.320249   3.645480   0.362 0.717323
## as.factor(inc)4             2.076024   3.722298   0.558 0.577179
## as.factor(inc)5             1.432838   3.767516   0.380 0.703808
## as.factor(inc)6             0.434202   4.032632   0.108 0.914281
## as.factor(inc)7             1.603763   3.731236   0.430 0.667436
## as.factor(inc)8             2.048847   5.447255   0.376 0.706919
## as.factor(inc)9            -2.597704   5.083590  -0.511 0.609486
## as.factor(smokef)smokes    -9.010719   1.180416  -7.634 6.18e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.71 on 843 degrees of freedom
## Multiple R-squared:  0.1681, Adjusted R-squared:  0.1434
## F-statistic: 6.814 on 25 and 843 DF,  p-value: < 2.2e-16
```

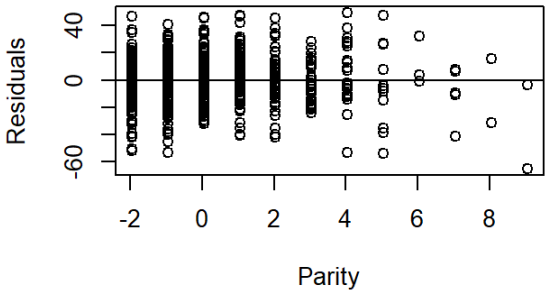
We're relatively satisfied with the plain vanilla model because residual plots don't show obvious patterns and dots are randomly scattered above and below x axis. Most predictors have pretty significant p-values even though r-square for this model is 0.1681. Next step, we try to improve fit of the model by adding in interaction effect.

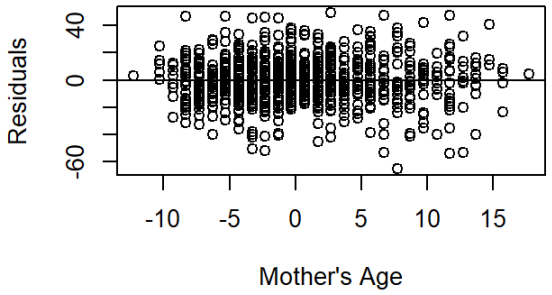
Second Model: interaction added

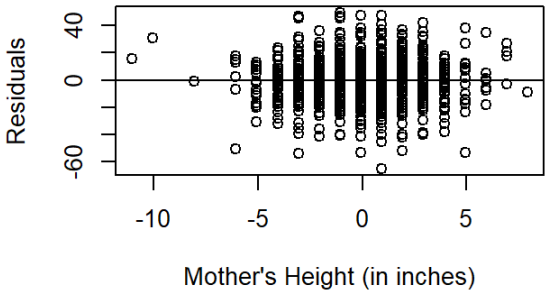
```
#second pass model
reg2 <- lm(bwt.oz ~ cdate + cparity + cmage + cmht * as.factor(mracef) + cmpregwt + as.factor(med) + as.fac
tor(inc) + as.factor(smoke), data = babies)
```

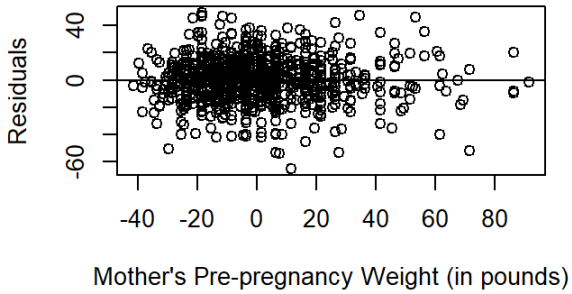
To do check of residuals versus each predictor to make sure validity of assumptions are satisfied:

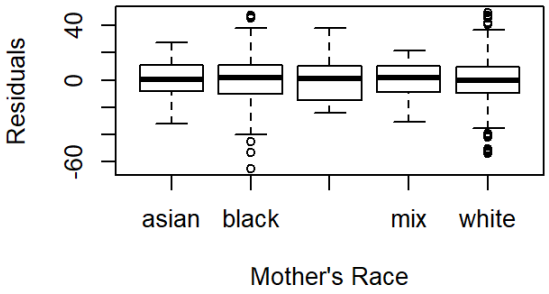

Residuals
40
0
-60
-200
-100
0
100
Birth Date

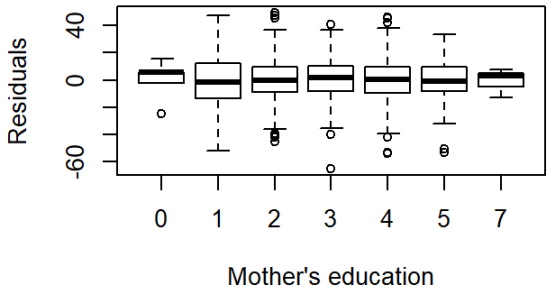

Residuals
40
0
-60
-2
0
2
4
6
8
Parity

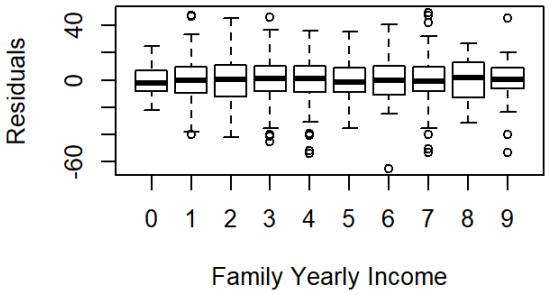

Residuals
40
0
-60
-10
-5
0
5
10
15
Mother's Age

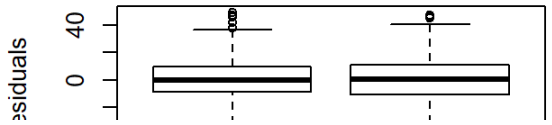

Residuals
40
0
-60
-10
-5
0
5
Mother's Height (in inches)

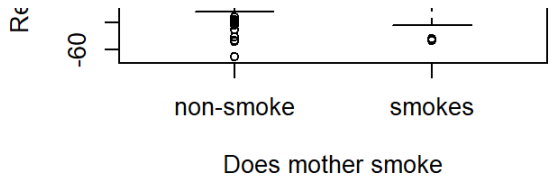

Residuals
40
0
-60
-40
-20
0
20
40
60
80
Mother's Pre-pregnancy Weight (in pounds)


Residuals
40
0
-60
asian
black
mix
white
Mother's Race


Residuals
40
0
-60
0
1
2
3
4
5
7
Mother's education


Residuals
40
0
-60
0
1
2
3
4
5
6
7
8
9
Family Yearly Income


esiduals
40
0



All plots look good (satisfied linearity and constant-variance assumption). Thus we look into its summary table:

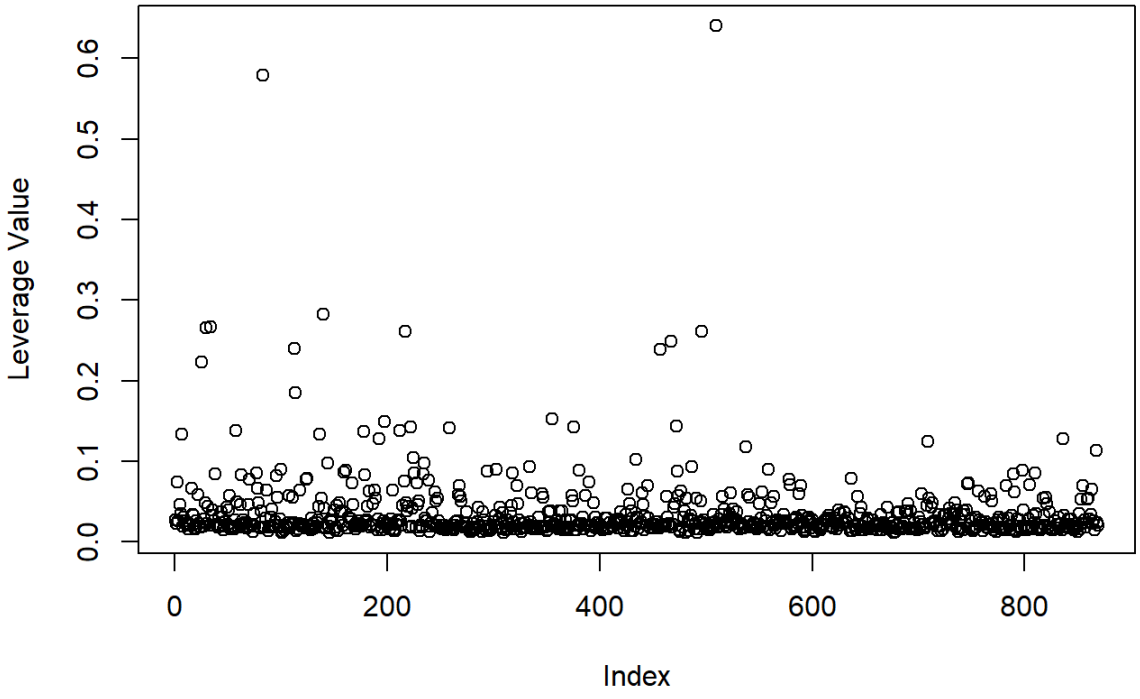
```
##
## Call:
## lm(formula = bwt.oz ~ cdate + cparity + cmage + cmht * as.factor(mracef) +
##     cmpregwt + as.factor(med) + as.factor(inc) + as.factor(smoke),
##     data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.506  -9.769  -0.007  10.411  49.476
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   89.468682  10.450899   8.561  < 2e-16 ***
## cdate                          0.012315   0.005434   2.266  0.02370 *
## cparity                        0.599741   0.404517   1.483  0.13855
## cmage                         -0.003600   0.133965  -0.027  0.97857
## cmht                          -3.753486   2.123932  -1.767  0.07755 .
## as.factor(mracef)black        10.412385   6.247476   1.667  0.09596 .
## as.factor(mracef)mexican      16.089097   7.583142   2.122  0.03416 *
## as.factor(mracef)mix          16.816305   7.554110   2.226  0.02627 *
## as.factor(mracef)white        19.350955   6.071329   3.187  0.00149 **
## cmpregwt                       0.103286   0.032816   3.147  0.00171 **
## as.factor(med)1               11.701503   8.252435   1.418  0.15658
## as.factor(med)2               14.110454   8.139362   1.734  0.08336 .
## as.factor(med)3               12.402640   8.413785   1.474  0.14083
## as.factor(med)4               14.462060   8.162968   1.772  0.07681 .
## as.factor(med)5               13.474569   8.196737   1.644  0.10057
## as.factor(med)7                1.422176  11.495993   0.124  0.90157
## as.factor(inc)1                3.144774   3.572864   0.880  0.37901
## as.factor(inc)2                4.852043   3.589642   1.352  0.17684
## as.factor(inc)3                1.274778   3.633965   0.351  0.72583
## as.factor(inc)4                2.188258   3.707896   0.590  0.55524
## as.factor(inc)5                1.669353   3.752835   0.445  0.65656
## as.factor(inc)6                0.659114   4.016069   0.164  0.86968
## as.factor(inc)7                1.697361   3.719657   0.456  0.64828
## as.factor(inc)8                1.926436   5.432107   0.355  0.72295
## as.factor(inc)9               -2.340019   5.068243  -0.462  0.64441
## as.factor(smoke)1             -9.241550   1.183652  -7.808 1.73e-14 ***
## cmht:as.factor(mracef)black    4.448936   2.183421   2.038  0.04190 *
## cmht:as.factor(mracef)mexican  1.265022   2.597743   0.487  0.62641
## cmht:as.factor(mracef)mix      5.358783   2.652647   2.020  0.04368 *
## cmht:as.factor(mracef)white    5.014934   2.135086   2.349  0.01906 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.63 on 839 degrees of freedom
## Multiple R-squared:  0.1797, Adjusted R-squared:  0.1513
## F-statistic: 6.336 on 29 and 839 DF,  p-value: < 2.2e-16
```

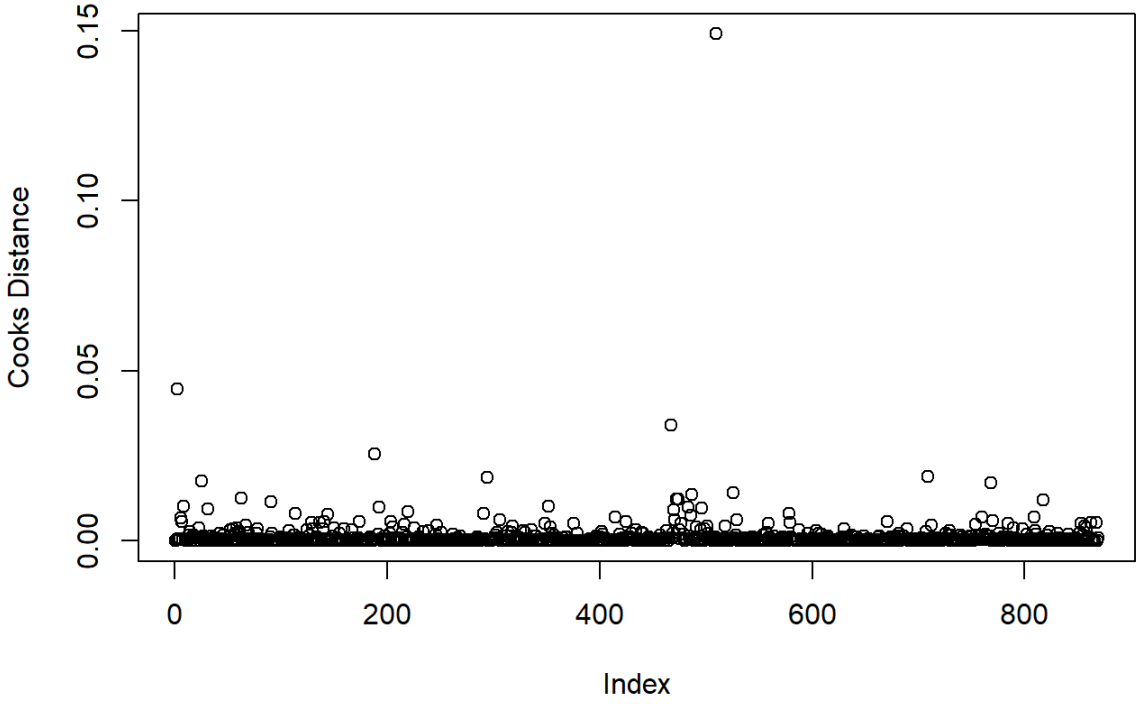
R-square becomes 0.1797 and fit of model is indeed improved.

Influence Diagnostics

Let's check case influence diagnostics based on the second model.

```
lev = hatvalues(reg2)
cooks = cooks.distance(reg2)
```





There is one point with cooks distance over 0.14, which is really far away fron others in terms of cooks distance. There are also two points far out there on leverge. Let's take a look at the person with largest cooks distance first.

```
babies[cooks > 0.14,]
```

```
##     date bwt.oz parity mrace mage med mht mpregwt inc smoke  mracef smokef
## 510 1713    146     10     6   39   0  53     110   3     1 mexican smokes
##        cdate  cparity    cmage      cmht  cmpregwt
## 510 176.5765 8.047181 11.70541 -11.06904 -18.47871
```

```
lev[510]
```

```
##       510
## 0.6402288
```

```
babies[lev > 0.5, ]
```

```
##     date bwt.oz parity mrace mage med mht mpregwt inc smoke  mracef
## 83  1553    105      3     9   22   2  56      93   3     0     mix
## 510 1713    146     10     6   39   0  53     110   3     1 mexican
##        smokef     cdate  cparity     cmage       cmht  cmpregwt
## 83  non-smoke  16.57652 1.047181 -5.294591  -8.069045 -35.47871
## 510    smokes 176.57652 8.047181 11.705409 -11.069045 -18.47871
```

The 510th mother with largest cook distance and a substantial leverage is really an exceptional data point. This mother with in total 10 previous pregnancies, lowest height (if you look in to the initial summary of this data set) and smoking habit has a above-average-weight baby, which is not what the model indicated. However, this data point wasn't caused by data entry error, so we can't delete it. We need to keep special cases.

Interpretations

Checking both residual plots and influence dianostics, we can now determine that out final model is reg2. Let's again take a look at the summary statistics and try to intrepret the coefficients:

```
##
## Call:
## lm(formula = bwt.oz ~ cdate + cparity + cmage + cmht * as.factor(mracef) +
##     cmpregwt + as.factor(med) + as.factor(inc) + as.factor(smoke),
##     data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.506  -9.769  -0.007  10.411  49.476
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  89.468682  10.450899   8.561  < 2e-16 ***
## cdate                         0.012315   0.005434   2.266  0.02370 *
## cparity                       0.599741   0.404517   1.483  0.13855
## cmage                        -0.003600   0.133965  -0.027  0.97857
## cmht                         -3.753486   2.123932  -1.767  0.07755 .
## as.factor(mracef)black       10.412385   6.247476   1.667  0.09596 .
## as.factor(mracef)mexican     16.089097   7.583142   2.122  0.03416 *
## as.factor(mracef)mix         16.816305   7.554110   2.226  0.02627 *
## as.factor(mracef)white       19.350955   6.071329   3.187  0.00149 **
## cmpregwt                      0.103286   0.032816   3.147  0.00171 **
## as.factor(med)1              11.701503   8.252435   1.418  0.15658
## as.factor(med)2              14.110454   8.139362   1.734  0.08336 .
## as.factor(med)3              12.402640   8.413785   1.474  0.14083
## as.factor(med)4              14.462060   8.162968   1.772  0.07681 .
## as.factor(med)5              13.474569   8.196737   1.644  0.10057
## as.factor(med)7               1.422176  11.495993   0.124  0.90157
## as.factor(inc)1               3.144774   3.572864   0.880  0.37901
## as.factor(inc)2               4.852043   3.589642   1.352  0.17684
## as.factor(inc)3               1.274778   3.633965   0.351  0.72583
## as.factor(inc)4               2.188258   3.707896   0.590  0.55524
## as.factor(inc)5               1.669353   3.752835   0.445  0.65656
## as.factor(inc)6               0.659114   4.016069   0.164  0.86968
## as.factor(inc)7               1.697361   3.719657   0.456  0.64828
## as.factor(inc)8               1.926436   5.432107   0.355  0.72295
## as.factor(inc)9              -2.340019   5.068243  -0.462  0.64441
## as.factor(smoke)1            -9.241550   1.183652  -7.808 1.73e-14 ***
## cmht:as.factor(mracef)black   4.448936   2.183421   2.038  0.04190 *
## cmht:as.factor(mracef)mexican 1.265022   2.597743   0.487  0.62641
## cmht:as.factor(mracef)mix     5.358783   2.652647   2.020  0.04368 *
## cmht:as.factor(mracef)white   5.014934   2.135086   2.349  0.01906 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.63 on 839 degrees of freedom
## Multiple R-squared:  0.1797, Adjusted R-squared:  0.1513
## F-statistic: 6.336 on 29 and 839 DF,  p-value: < 2.2e-16
```

```
##                                       2.5 %       97.5 %
## (Intercept)                     68.955704419 109.98166036
## cdate                            0.001648246   0.02298084
## cparity                         -0.194242185   1.39372513
## cmage                           -0.266544934   0.25934580
## cmht                            -7.922330397   0.41535853
## as.factor(mracef)black          -1.850133552  22.67490271
## as.factor(mracef)mexican         1.204940409  30.97325357
## as.factor(mracef)mix             1.989130823  31.64347836
## as.factor(mracef)white           7.434177535  31.26773244
## cmpregwt                         0.038874746   0.16769755
## as.factor(med)1                 -4.496339287  27.89934494
## as.factor(med)2                 -1.865450392  30.08635764
## as.factor(med)3                 -4.111900252  28.91717999
## as.factor(med)4                 -1.560177450  30.48429815
## as.factor(med)5                 -2.613949083  29.56308725
## as.factor(med)7                -21.142107184  23.98646010
## as.factor(inc)1                 -3.868028728  10.15757596
## as.factor(inc)2                 -2.193689148  11.89777574
## as.factor(inc)3                 -5.857951585   8.40750786
## as.factor(inc)4                 -5.089583262   9.46609914
## as.factor(inc)5                 -5.696694920   9.03540095
## as.factor(inc)6                 -7.223607902   8.54183510
## as.factor(inc)7                 -5.603565406   8.99828751
## as.factor(inc)8                 -8.735679285  12.58855227
## as.factor(inc)9                -12.287942543   7.60790544
## as.factor(smoke)1              -11.564817193  -6.91828251
## cmht:as.factor(mracef)black      0.163326212   8.73454585
## cmht:as.factor(mracef)mexican   -3.833816115   6.36386058
## cmht:as.factor(mracef)mix        0.152180314  10.56538664
## cmht:as.factor(mracef)white      0.824197120   9.20567025
```

R-square of the model is 0.1797, which indicates that we probably need to add better predictors to reach a better fit. But now this is the best we can get with current data set.

Intercept Mean weight of a baby, born on the 441st day counting from January 1, 1961, given birth by an asian non-smoking non-education mother with mean parity, age, heigh, pre-pregnancy weight, who comes from a mean-income family, is 89.468682 oz (95% CI:68.9557, 109.9817).

Parity (and all other continuous variables could be interpreted the same way) If we keep all other predictors constant and increase parity by 1, mean weight of baby will increase by 0.5997 oz (95% CI:-0.1942, 1.3937)

Mother's Height If we keep all other predictors constant and increase mother's height by 1 inch, mean weight of baby will increase by -3.753 oz (95% CI: 7.9223, 0.4154).

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 Asian to Black, mean weight of baby will increase by 10.4124 oz (95% CI: -1.8501, 22.6749).

Interaction between Mother's Height and Mother's Race If we keep all other predictors constant, we change both mother's race from Asian to Black and increase mother's height by 1 inch, then mean weight of baby will increase $ 10.4124+4.4489 = 14.8613 $ oz. (95% CI:$-1.8501 + 0.1633, 22.6749 + 8.7345$)

Nested F test to see if smoke is a useful predictor (Question 1)

```
## Analysis of Variance Table
##
## Model 1: bwt.oz ~ cdate + cparity + cmage + cmht * as.factor(mracef) +
##     cmpregwt + as.factor(med) + as.factor(inc)
## Model 2: bwt.oz ~ cdate + cparity + cmage + cmht * as.factor(mracef) +
##     cmpregwt + as.factor(med) + as.factor(inc) + as.factor(smoke)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    840 248865
## 2    839 232008  1     16857 60.959 1.735e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Smoke appears to be a useful predictor because p-value for the nested f test is 1.735e-14, which is really small and below 0.05.

What is a likely range for the difference in birth weights for smokers and non-smokers?

```
##                                       2.5 %       97.5 %
## (Intercept)                     68.955704419 109.98166036
## cdate                            0.001648246   0.02298084
## cparity                         -0.194242185   1.39372513
## cmage                           -0.266544934   0.25934580
## cmht                            -7.922330397   0.41535853
## as.factor(mracef)black          -1.850133552  22.67490271
## as.factor(mracef)mexican         1.204940409  30.97325357
## as.factor(mracef)mix             1.989130823  31.64347836
## as.factor(mracef)white           7.434177535  31.26773244
## cmpregwt                         0.038874746   0.16769755
## as.factor(med)1                 -4.496339287  27.89934494
## as.factor(med)2                 -1.865450392  30.08635764
## as.factor(med)3                 -4.111900252  28.91717999
## as.factor(med)4                 -1.560177450  30.48429815
## as.factor(med)5                 -2.613949083  29.56308725
## as.factor(med)7                -21.142107184  23.98646010
## as.factor(inc)1                 -3.868028728  10.15757596
## as.factor(inc)2                 -2.193689148  11.89777574
## as.factor(inc)3                 -5.857951585   8.40750786
## as.factor(inc)4                 -5.089583262   9.46609914
## as.factor(inc)5                 -5.696694920   9.03540095
## as.factor(inc)6                 -7.223607902   8.54183510
## as.factor(inc)7                 -5.603565406   8.99828751
## as.factor(inc)8                 -8.735679285  12.58855227
## as.factor(inc)9                -12.287942543   7.60790544
## as.factor(smoke)1              -11.564817193  -6.91828251
## cmht:as.factor(mracef)black      0.163326212   8.73454585
## cmht:as.factor(mracef)mexican   -3.833816115   6.36386058
## cmht:as.factor(mracef)mix        0.152180314  10.56538664
## cmht:as.factor(mracef)white      0.824197120   9.20567025
```

From the confidence interval summary, we're 95% confident that mother who smokes tends to give birth to a baby weigh 11.5648 to 6.9183 less than mother who never smokes holding all other variables constant.

Nested F test to see if the interaction between birth weight and mother's race is significant (Question 2)

```
## Analysis of Variance Table
##
## Model 1: bwt.oz ~ cdate + cparity + cmage + cmht + cmpregwt + as.factor(mracef) +
##     as.factor(med) + as.factor(inc) + as.factor(smokef)
## Model 2: bwt.oz ~ cdate + cparity + cmage + cmht + cmpregwt + as.factor(med) +
##     as.factor(inc) + as.factor(smokef) * as.factor(mracef)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    843 235278
## 2    839 233460  4    1818.2 1.6336 0.1637
```

P-value for the nested f test between model with interaction and plain vanilla model is 0.1637 which is bigger than 0.05 for sure. Thus, there is no evidence that the association between smoking and birth weight differs by mother's race.

Are there other intersting associations with birth weight that are worth mentioning? (Question 3)

Not necessarily, I've already explored association between mother's height and birth weight differs by mother's race.