Homework3

Xuan Yu

9/24/2018

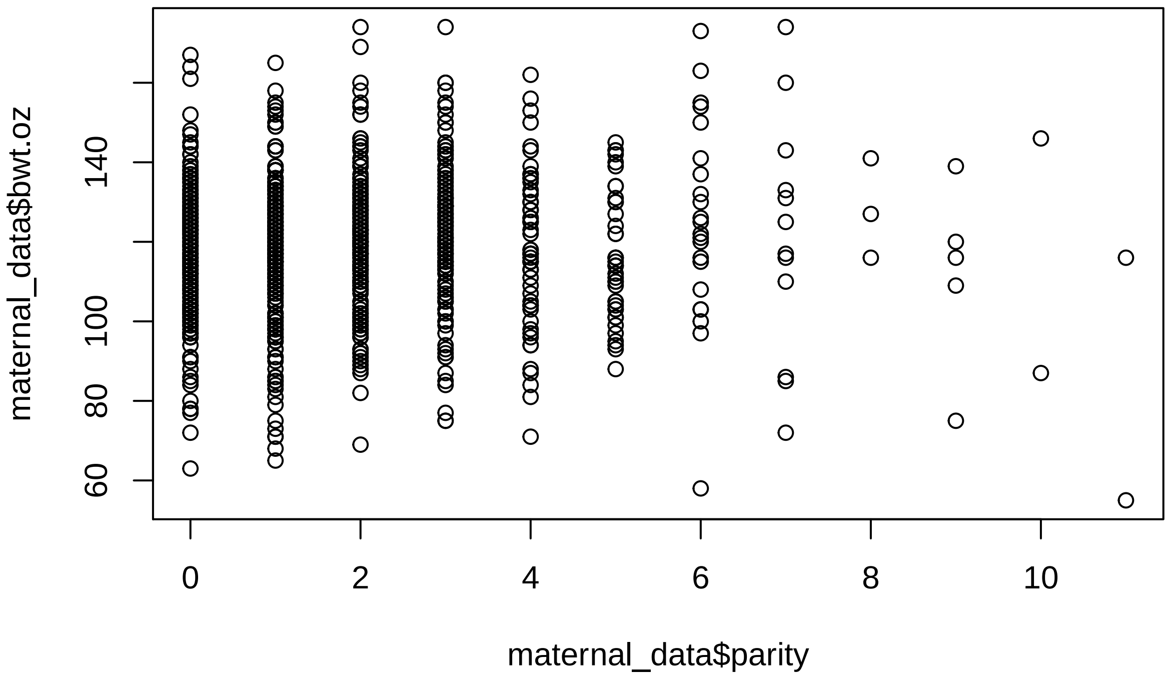
Maternal Smoking and Birth Weights

Read in the data.

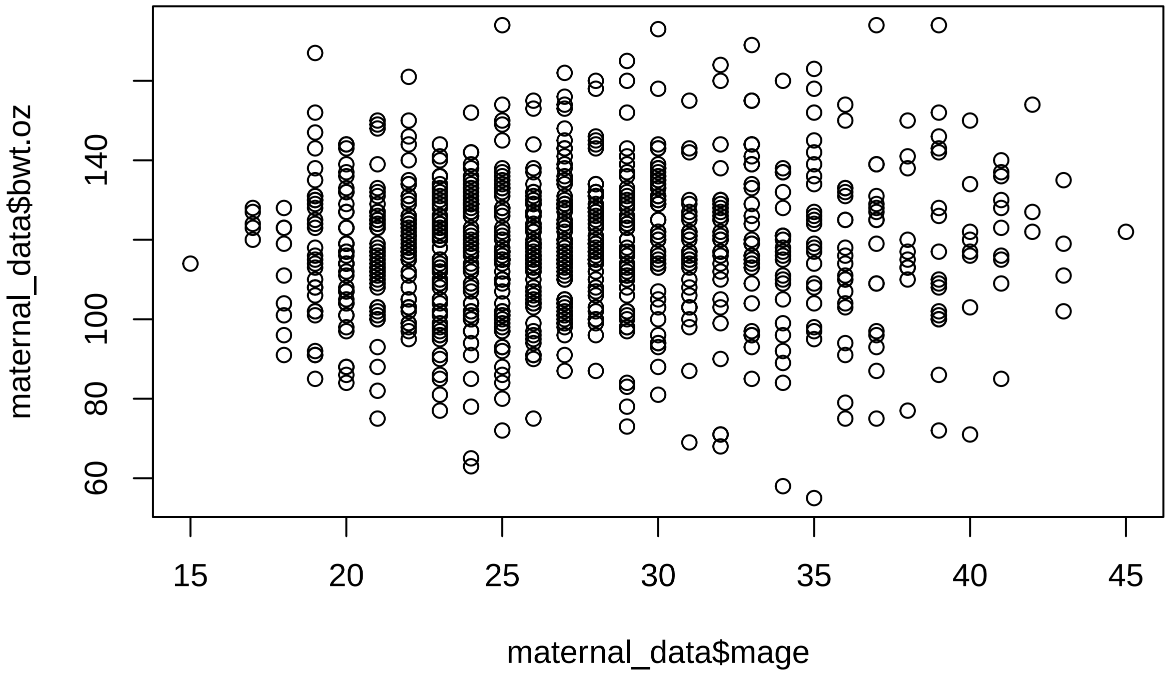
```
maternal_data <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW3/smoking.csv")
maternal_data$mrace_new <- maternal_data$mrace
maternal_data$mrace_new[maternal_data$mrace >= 0 & maternal_data$mrace <= 5] <- 5

maternal_data$who_smoke <- maternal_data$smoke
maternal_data$who_smoke[maternal_data$smoke != 0] <- 1
```

Plot all the continuous variable and categorical variable:

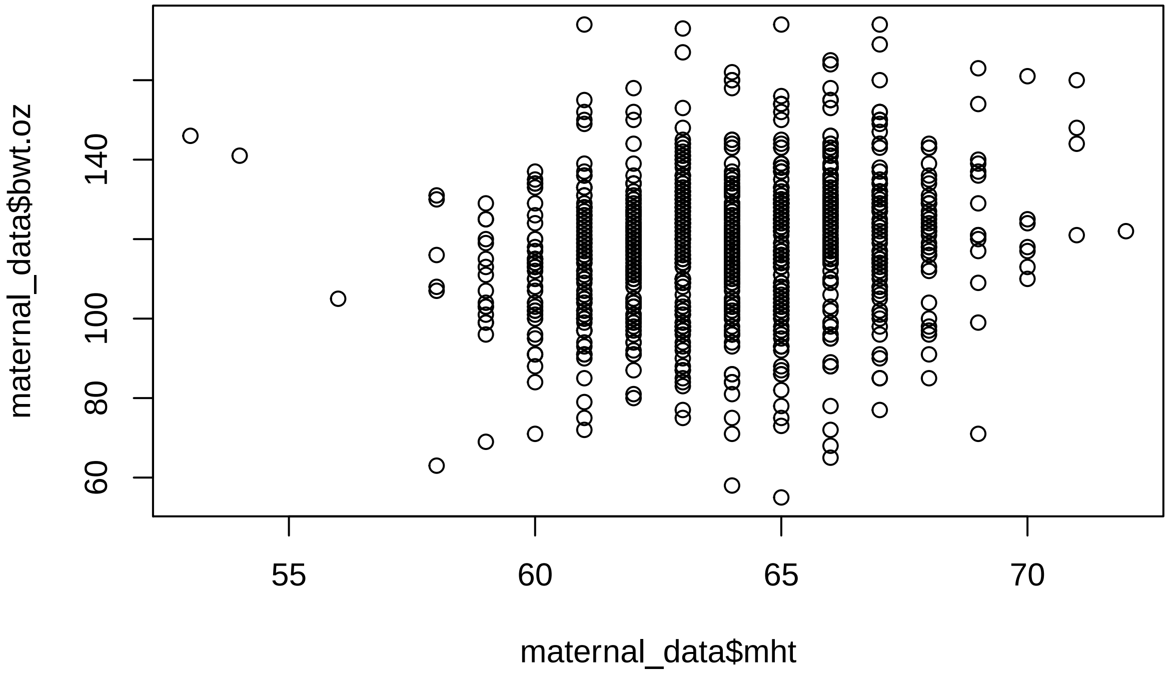
```
plot(maternal_data$parity, maternal_data$bwt.oz)
```



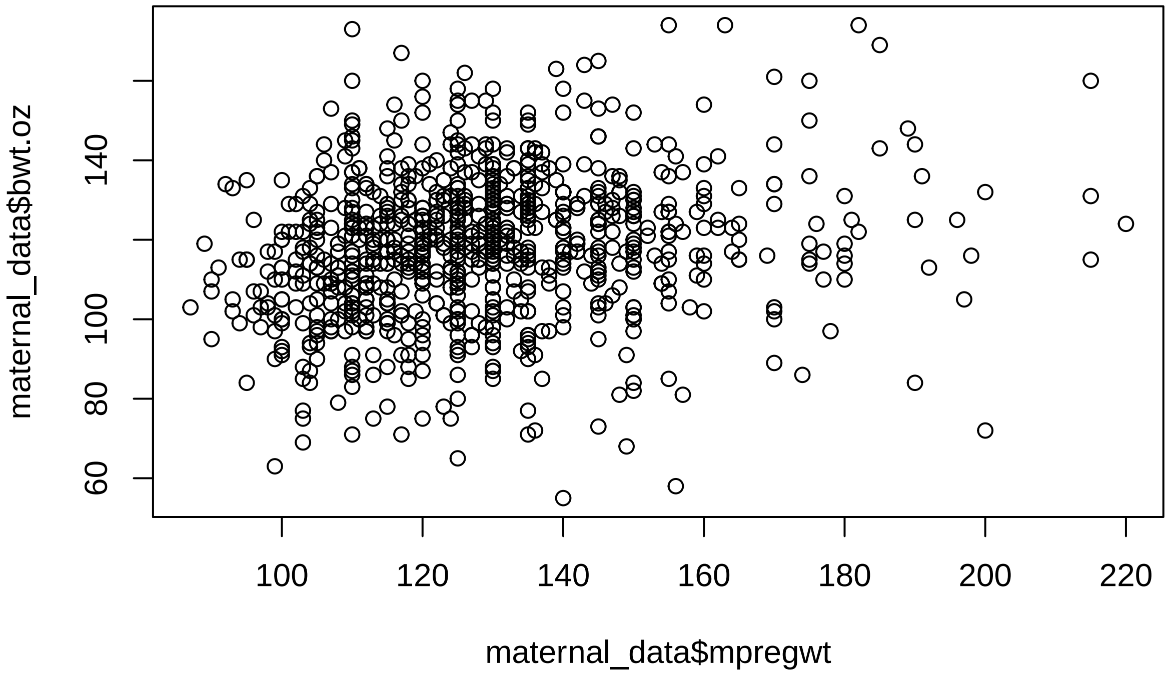
```
plot(maternal_data$mage, maternal_data$bwt.oz)
```



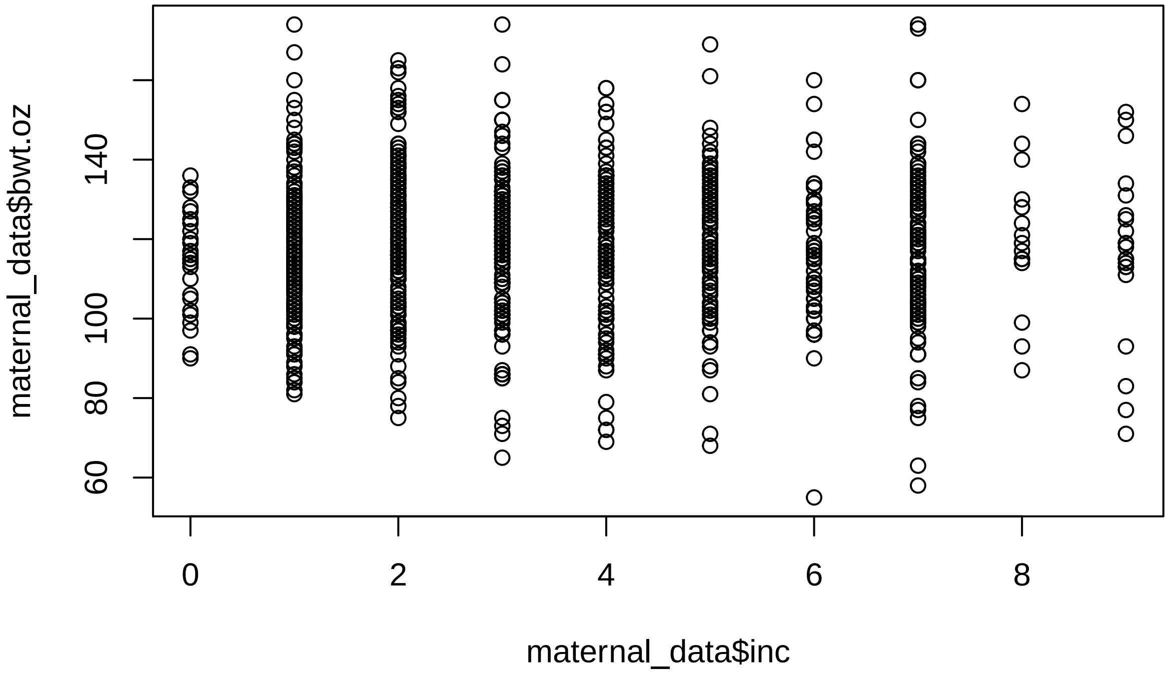
```
plot(maternal_data$mht, maternal_data$bwt.oz)
```



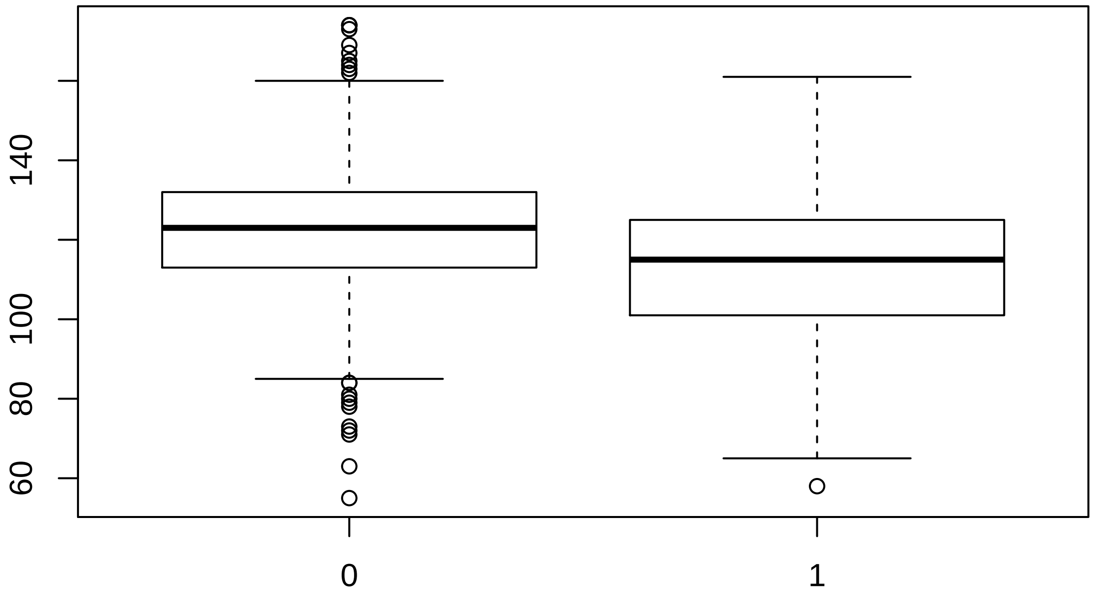
```
plot(maternal_data$mpregwt, maternal_data$bwt.oz)
```



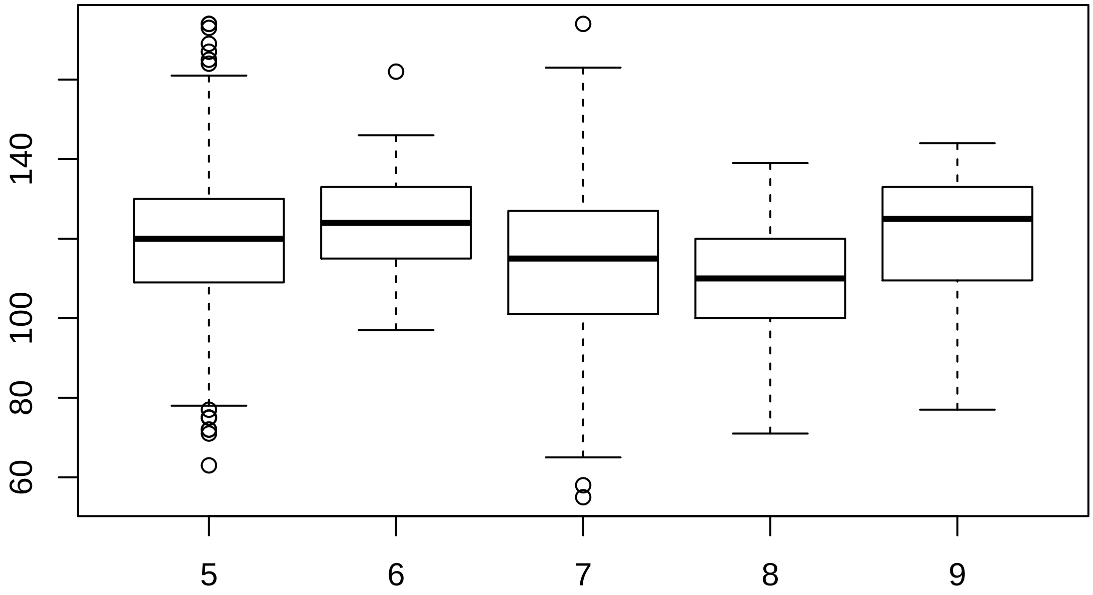
```r
plot(maternal_data$inc, maternal_data$bwt.oz)
```



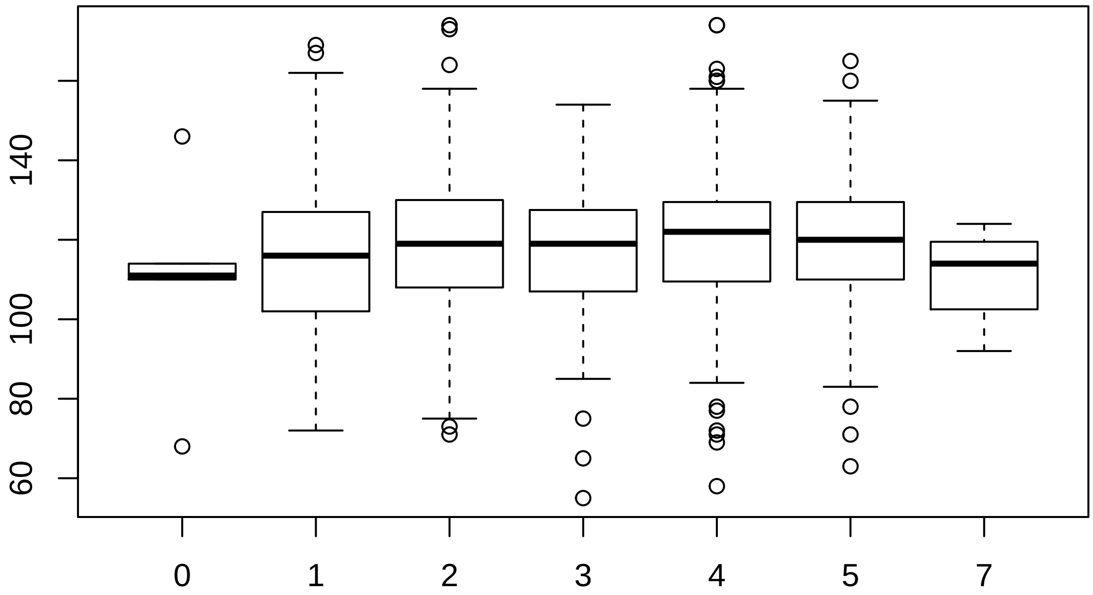
```r
boxplot(bwt.oz~who_smoke, data = maternal_data)
```



```r
boxplot(bwt.oz~as.factor(mrace_new), data = maternal_data)
```



```r
boxplot(bwt.oz~as.factor(med), data = maternal_data)
```



See if there are correlations between predictor variables, there's no very huge correlations:

```r
round(cor(maternal_data[,4:14]), 3)
```

```
##            bwt.oz parity  mrace   mage    med    mht mpregwt    inc  smoke
## bwt.oz      1.000  0.041 -0.130  0.044  0.038  0.188   0.182  0.002 -0.249
## parity      0.041  1.000  0.149  0.524 -0.201 -0.043   0.151  0.009  0.011
## mrace      -0.130  0.149  1.000  0.014 -0.079 -0.165   0.023 -0.122 -0.114
## mage        0.044  0.524  0.014  1.000  0.134 -0.005   0.146  0.297 -0.070
## med         0.038 -0.201 -0.079  0.134  1.000  0.115  -0.054  0.217 -0.138
## mht         0.188 -0.043 -0.165 -0.005  0.115  1.000   0.460  0.071  0.041
## mpregwt     0.182  0.151  0.023  0.146 -0.054  0.460   1.000 -0.005 -0.049
## inc         0.002  0.009 -0.122  0.297  0.217  0.071  -0.005  1.000  0.007
## smoke      -0.249  0.011 -0.114 -0.070 -0.138  0.041  -0.049  0.007  1.000
## mrace_new -0.149  0.156  0.822  0.038 -0.019 -0.151   0.039 -0.113 -0.108
## who_smoke -0.249  0.011 -0.114 -0.070 -0.138  0.041  -0.049  0.007  1.000
##            mrace_new who_smoke
## bwt.oz        -0.149    -0.249
## parity         0.156     0.011
## mrace          0.822    -0.114
## mage           0.038    -0.070
## med           -0.019    -0.138
## mht           -0.151     0.041
## mpregwt        0.039    -0.049
## inc           -0.113     0.007
## smoke         -0.108     1.000
## mrace_new      1.000    -0.108
## who_smoke     -0.108     1.000
```

Maternal height and weight variable have quadratic trends, so take a quadratic transformation for maternal height and weight, then do the modeling:

```r
maternal_data$mht2 <- maternal_data$mht ^ 2
maternal_data$mpregwt2 <- maternal_data$mpregwt ^ 2

maternal_lm <- lm(bwt.oz ~ date + who_smoke + parity + mage +
                  mht + mht2 + mpregwt + mpregwt2 + inc +
                  as.factor(mrace_new) + as.factor(med),
                  data = maternal_data)
summary(maternal_lm)
```

```
## 
## Call:
## lm(formula = bwt.oz ~ date + who_smoke + parity + mage + mht + 
##     mht2 + mpregwt + mpregwt2 + inc + as.factor(mrace_new) + 
##     as.factor(med), data = maternal_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -68.064  -9.536  -0.191  10.212  50.405 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            4.080e+02  2.469e+02   1.652   0.0989 .  
## date                   1.246e-02  5.398e-03   2.308   0.0212 *  
## who_smoke             -8.993e+00  1.176e+00  -7.645 5.64e-14 ***
## parity                 8.109e-01  3.962e-01   2.047   0.0410 *  
## mage                  -5.953e-02  1.333e-01  -0.447   0.6552    
## mht                   -1.196e+01  7.794e+00  -1.535   0.1252    
## mht2                   1.002e-01  6.095e-02   1.644   0.1005    
## mpregwt                5.396e-01  2.472e-01   2.183   0.0293 *  
## mpregwt2              -1.528e-03  8.621e-04  -1.773   0.0767 .  
## inc                   -3.933e-01  2.741e-01  -1.435   0.1517    
## as.factor(mrace_new)6  3.490e+00  3.495e+00   0.999   0.3183    
## as.factor(mrace_new)7 -9.240e+00  1.564e+00  -5.909 4.97e-09 ***
## as.factor(mrace_new)8 -6.845e+00  3.108e+00  -2.203   0.0279 *  
## as.factor(mrace_new)9 -2.828e+00  4.400e+00  -0.643   0.5205    
## as.factor(med)1        8.207e+00  7.860e+00   1.044   0.2967    
## as.factor(med)2        1.048e+01  7.739e+00   1.354   0.1761    
## as.factor(med)3        8.984e+00  8.035e+00   1.118   0.2639    
## as.factor(med)4        1.091e+01  7.773e+00   1.404   0.1607    
## as.factor(med)5        9.866e+00  7.808e+00   1.264   0.2067    
## as.factor(med)7       -1.408e+00  1.130e+01  -0.125   0.9009    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 16.64 on 849 degrees of freedom
## Multiple R-squared:  0.1689, Adjusted R-squared:  0.1503 
## F-statistic: 9.079 on 19 and 849 DF,  p-value: < 2.2e-16
```

We checked the assumptions and all the assumptions are met.

For question 2, We do the regression with interaction of race variable and do the nested F test. The result p value is 0.3079, so we don't find interaction of race variable significant:

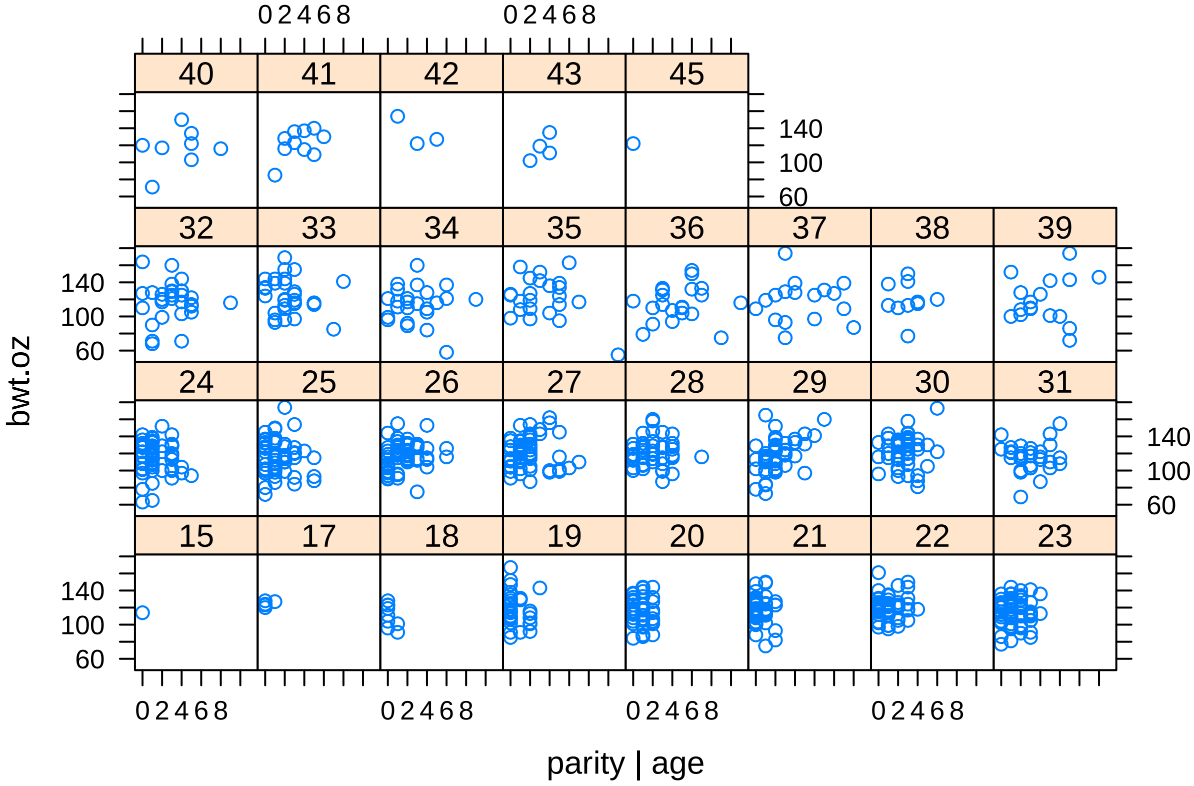
```
maternal_lm_race_interaction <- lm(bwt.oz ~ date + who_smoke * as.factor(mrace_new) +
                                   parity + mage + mht + mht2 + mpregwt + mpregwt2 +
                                   inc + as.factor(mrace_new) + as.factor(med),
                                   data = maternal_data)
anova(maternal_lm, maternal_lm_race_interaction)
```

```
## Analysis of Variance Table
```

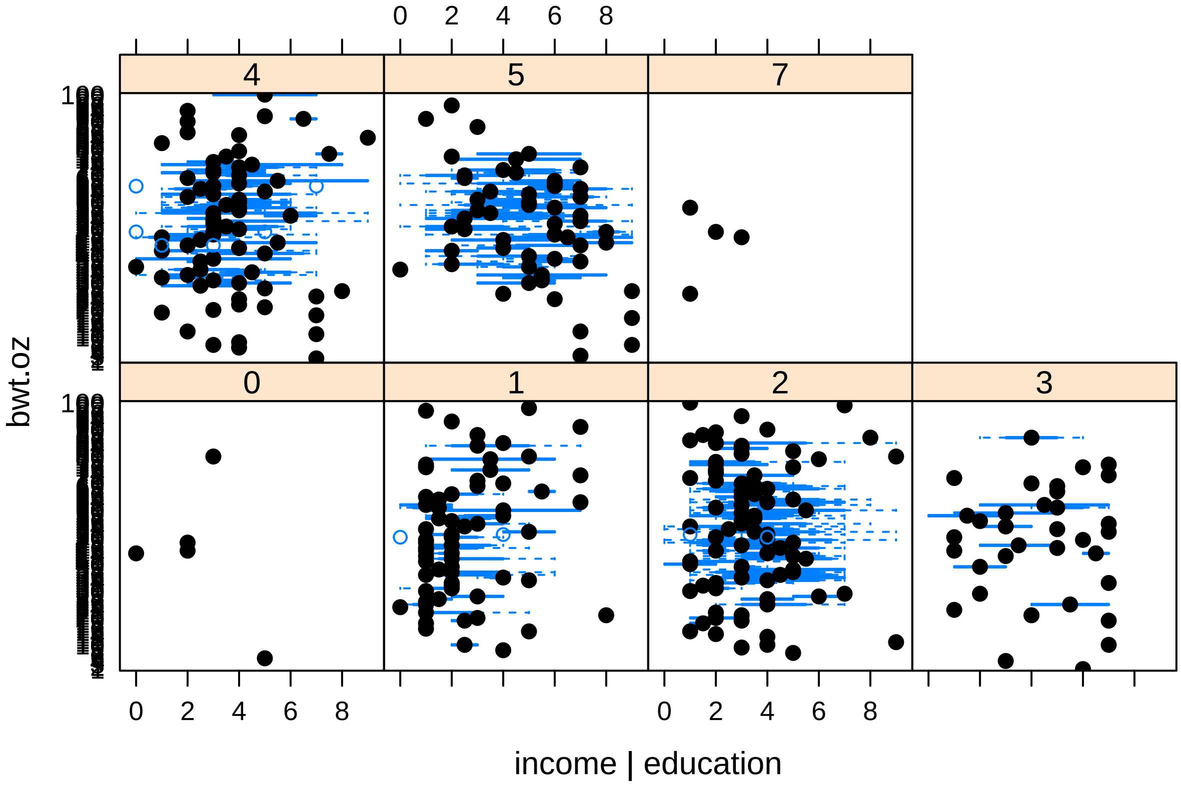
```
##
## Model 1: bwt.oz ~ date + who_smoke + parity + mage + mht + mht2 + mpregwt +
##     mpregwt2 + inc + as.factor(mrace_new) + as.factor(med)
## Model 2: bwt.oz ~ date + who_smoke * as.factor(mrace_new) + parity + mage +
##     mht + mht2 + mpregwt + mpregwt2 + inc + as.factor(mrace_new) +
##     as.factor(med)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    849 235061
## 2    845 233730  4    1331.2 1.2032 0.3079
```

We also need to check if there are other interactions, and found parity and age, as well as education and income might have interaction between each other:

```
library("lattice")
xyplot(bwt.oz~parity|as.factor(mage), data = maternal_data, xlab = "parity | age")
```



```
bwplot(bwt.oz~inc|as.factor(med), data = maternal_data, xlab = "income | education")
```



We add these two interections to our final model:

```
maternal_lm_med_interaction <- lm(bwt.oz ~ date + who_smoke + parity * mage + mage +
                                  mht + mht2 + mpregwt + mpregwt2 + inc +
                                  as.factor(mrace_new) + as.factor(med) * inc,
                                  data = maternal_data)
summary(maternal_lm_med_interaction)
```

```
##
## Call:
## lm(formula = bwt.oz ~ date + who_smoke + parity * mage + mage +
##     mht + mht2 + mpregwt + mpregwt2 + inc + as.factor(mrace_new) +
##     as.factor(med) * inc, data = maternal_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.885 -10.154  -0.031   9.939  51.014
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.557e+02  2.474e+02   1.842   0.0658 .
## date                    1.258e-02  5.384e-03   2.337   0.0197 *
## who_smoke              -9.219e+00  1.179e+00  -7.820 1.58e-14 ***
## parity                  3.544e+00  1.721e+00   2.059   0.0398 *
## mage                    6.672e-02  1.645e-01   0.406   0.6851
## mht                    -1.327e+01  7.782e+00  -1.705   0.0886 .
## mht2                    1.098e-01  6.085e-02   1.805   0.0714 .
## mpregwt                 5.611e-01  2.467e-01   2.274   0.0232 *
## mpregwt2               -1.592e-03  8.605e-04  -1.849   0.0647 .
## inc                   -3.733e+00  4.598e+00  -0.812   0.4171
## as.factor(mrace_new)6  4.241e+00  3.512e+00   1.208   0.2276
## as.factor(mrace_new)7 -8.819e+00  1.572e+00  -5.609 2.76e-08 ***
## as.factor(mrace_new)8 -7.456e+00  3.110e+00  -2.398   0.0167 *
## as.factor(mrace_new)9 -2.972e+00  4.402e+00  -0.675   0.4999
## as.factor(med)1       -5.345e+00  1.370e+01  -0.390   0.6964
## as.factor(med)2        8.655e-01  1.352e+01   0.064   0.9490
## as.factor(med)3       -7.870e-01  1.440e+01  -0.055   0.9564
## as.factor(med)4       -1.869e-01  1.360e+01  -0.014   0.9890
## as.factor(med)5        6.681e+00  1.366e+01   0.489   0.6250
## as.factor(med)7       -1.502e+01  2.374e+01  -0.633   0.5270
## parity:mage           -8.871e-02  5.417e-02  -1.638   0.1018
## inc:as.factor(med)1    4.999e+00  4.680e+00   1.068   0.2858
## inc:as.factor(med)2    3.452e+00  4.624e+00   0.746   0.4556
## inc:as.factor(med)3    3.546e+00  4.733e+00   0.749   0.4540
## inc:as.factor(med)4    3.860e+00  4.636e+00   0.833   0.4054
## inc:as.factor(med)5    2.097e+00  4.631e+00   0.453   0.6508
## inc:as.factor(med)7    5.271e+00  1.111e+01   0.474   0.6354
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.58 on 842 degrees of freedom
## Multiple R-squared:  0.1818, Adjusted R-squared:  0.1566
## F-statistic: 7.197 on 26 and 842 DF,  p-value: < 2.2e-16
```

Here is the 95% confindent interval:

```
confint(maternal_lm_med_interaction)
```

```
##                               2.5 %        97.5 %
## (Intercept)           -29.813871509  9.412272e+02
## date                    0.002013495  2.315017e-02
## who_smoke             -11.532828687 -6.905031e+00
## parity                  0.165001754  6.922168e+00
## mage                   -0.256084433  3.895184e-01
## mht                   -28.542676543  2.007461e+00
## mht2                   -0.009578766  2.292785e-01
## mpregwt                 0.076845355  1.045383e+00
## mpregwt2               -0.003280449  9.749242e-05
## inc                   -12.757904001  5.291860e+00
## as.factor(mrace_new)6  -2.652109307  1.113335e+01
## as.factor(mrace_new)7 -11.905650934 -5.733269e+00
## as.factor(mrace_new)8 -13.559880025 -1.353031e+00
## as.factor(mrace_new)9 -11.612437436  5.669165e+00
## as.factor(med)1       -32.230008222  2.153917e+01
## as.factor(med)2       -25.662572747  2.739351e+01
## as.factor(med)3       -29.055565378  2.748161e+01
## as.factor(med)4       -26.878117676  2.650431e+01
## as.factor(med)5       -20.139448284  3.350083e+01
## as.factor(med)7       -61.617437408  3.156941e+01
## parity:mage            -0.195031067  1.760570e-02
## inc:as.factor(med)1    -4.187508694  1.418585e+01
## inc:as.factor(med)2    -5.624026056  1.252756e+01
## inc:as.factor(med)3    -5.743795751  1.283487e+01
## inc:as.factor(med)4    -5.240413648  1.296024e+01
## inc:as.factor(med)5    -6.991897018  1.118612e+01
## inc:as.factor(med)7   -16.537421947  2.707889e+01
```

Interpretation:

Answer for question 1:

Holding other variable constant, mothers who smoke tend to give birth to babies with 9.219 ounces lower weights than mothers who do not smoke. The 95% confidence interval for the difference in birth weights for smokers and non-smokers is (-11.5328, -6.9050) when considering non-smoker variable the base case.

Answer for question 2:

We did the nested F test for model with and without interaction of race, and found the result p value is $0.3079 > 0.05$, so we don't find the interaction of race variable important to smoking variable.

Answer for question 3:

1, We found it interesting that age seems to have interaction with parity, which makes sense: older mother tend to have larger number of pregnancies.

2, We found another interesting association between income and education level, that is, the association between income and birth weight differs by education level.