440 Case Study I

Jake Epstein, Daniel Spottiswood, Michael Tan, Sahil Patel, Man-Lin Hsiao

9/3/2019

Set Up

Load Necessary Packages

```
## load packages
library(dplyr)
library(ggplot2)
library(MASS)
library(gridExtra)
knitr::opts_chunk$set(warning=FALSE)
```

Load and Clean Data

```
## read in data
births = read.csv("data/Yr1116Birth.csv", na.strings = "9999")
deaths = read.csv("data/Yr1116Death.csv")

## rewrite NAs
births$SEX[which(births$SEX == 9)] = NA
births$CIGPN[which(births$CIGPN == 99)] = NA
births$CIGFN[which(births$CIGFN == 99)] = NA
births$CIGSN[which(births$CIGSN == 99)] = NA
births$CIGLN[which(births$CIGLN == 99)] = NA
births$PARITY[which(births$PARITY == 99)] = NA
births$PLUR[which(births$PLUR == 99)] = NA
births$GEST[which(births$GEST == 99)] = NA
births$MAGE[which(births$MAGE == 99)] = NA
select = dplyr::select
```

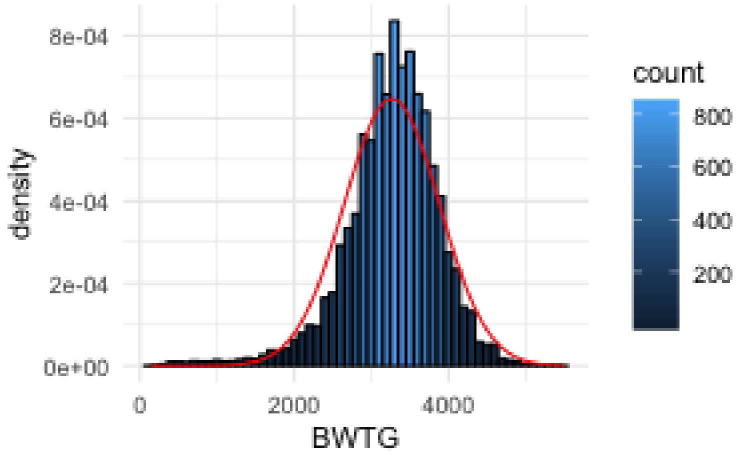
```
births = sample_n(births, 10000)
deaths = sample_n(deaths, 1000)
```

Exploratory Data Analysis

Birthweight

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     138    2948    3317    3264    3657    5505       8
```

```
## [1] 616.4405
```



Birthweight is close to normally distrubted, with a slight left skew, centered around ~3300g with a standard deviation of 600g. There appear to be no large outliers in terms of birthweight. 430 birth weights are missing. We see that the left tail is much larger than we would expect in a normal distribution

Smoking

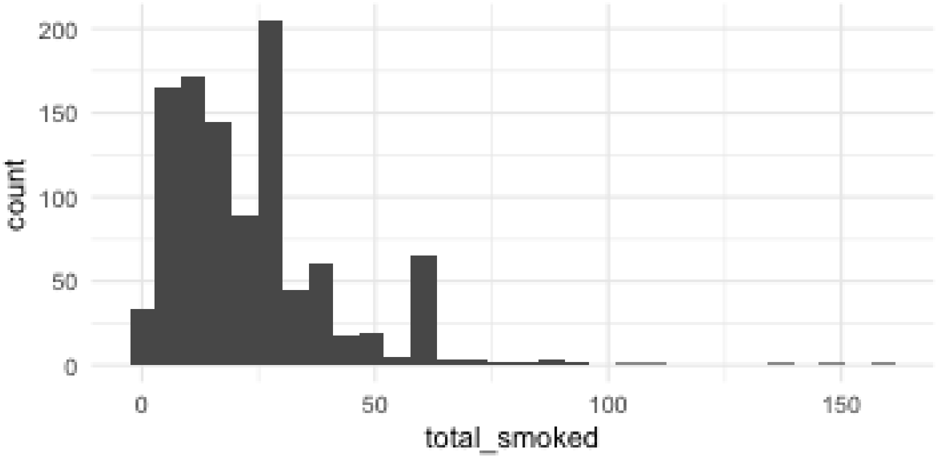
```
## [1] 0.1044237
```

```
## [1] 0.1366235
```

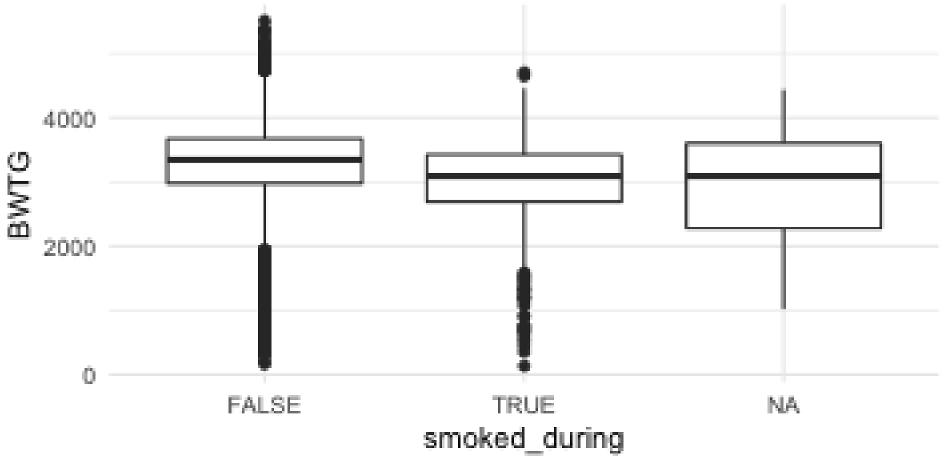
```
## [1] 0.1010131
```

```
## [1] 0.08564838
```

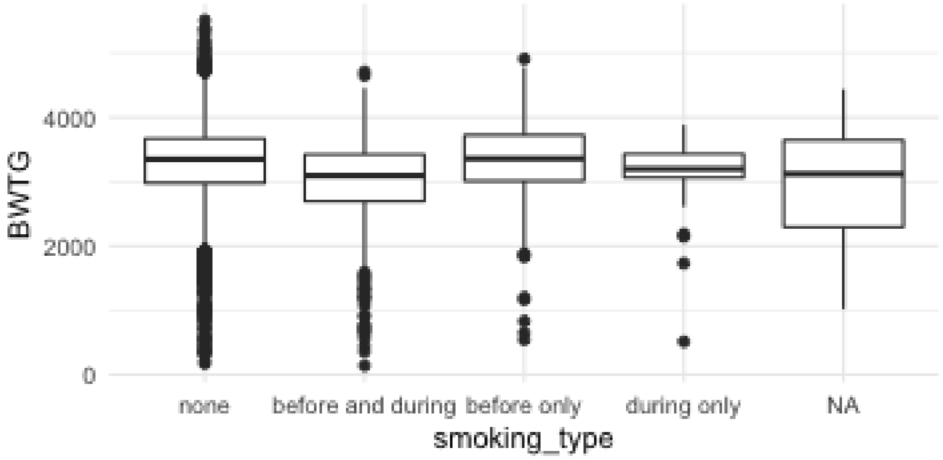
```
## [1] 0.08265623
```



```
## [1] 23.61575
```



```
##
## Call:
## lm(formula = BWTG ~ smoked_during, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3124.28  -317.28    50.72   383.96  2210.72
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3294.283      6.453  510.51   <2e-16 ***
## smoked_duringTRUE -276.238     19.973  -13.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 609.6 on 9961 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.01884,    Adjusted R-squared:  0.01874
## F-statistic: 191.3 on 1 and 9961 DF,  p-value: < 2.2e-16
```



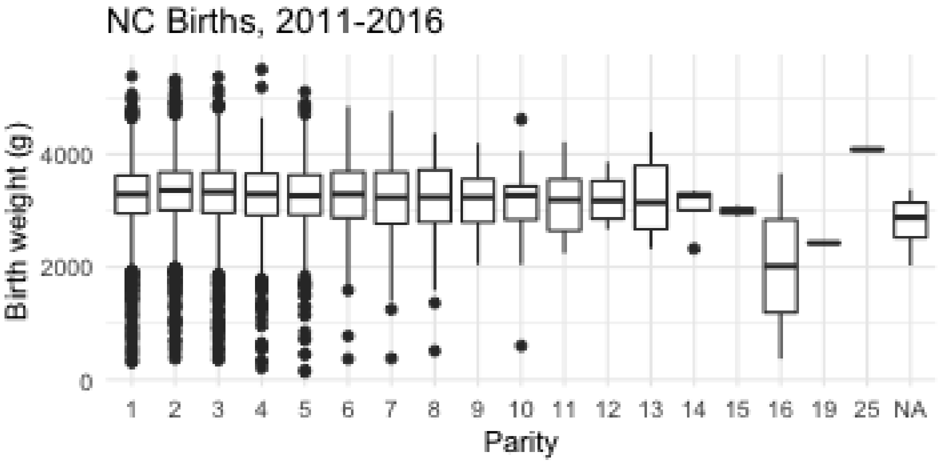
```
##
## Call:
## lm(formula = BWTG ~ smoking_type, data = births)
```

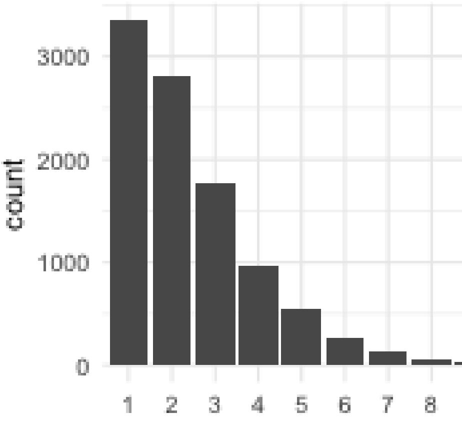
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3122.88  -315.88    52.12   386.53  2212.12
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3292.879      6.585 500.076   <2e-16 ***
## smoking_typebefore and during -277.409     20.307 -13.661   <2e-16 ***
## smoking_typebefore only         34.388     33.151   1.037   0.2996
## smoking_typeduring only       -196.273    106.318  -1.846   0.0649 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 609.6 on 9958 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.019,  Adjusted R-squared:  0.0187
## F-statistic: 64.28 on 3 and 9958 DF,  p-value: < 2.2e-16
```

Around 13% of women smoked in the three months leading up to pregnancy and around 10% of women at any point during their pregnancy. Among those who did smoke during pregnance, the average number of cigarettes smoked during pregnancy was 23. The birthweight of children of smokers was significantly lower than that of the children of nonsmokers, with an average difference of 231 grams. There is also a significant relationship between birthweight and smoking before pregnancy, even for those who did not smoke during pregnancy.

Parity

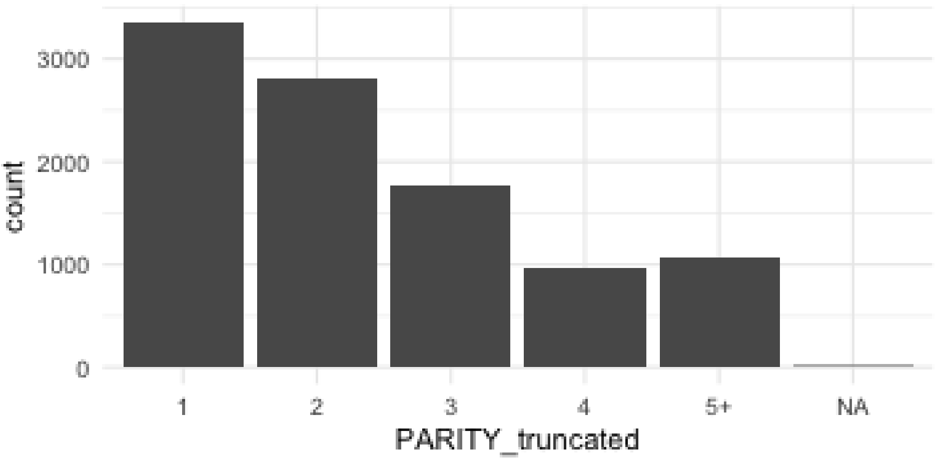
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   2.000   2.477   3.000  25.000      11
```

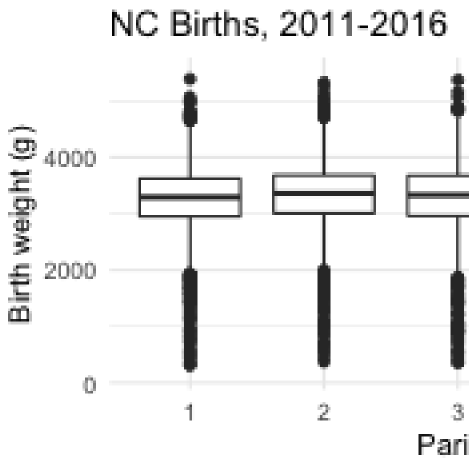




```
##
## Call:
## lm(formula = BWTG ~ PARITY, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3100.03  -313.57    47.07   384.67  2260.42
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3247.932     10.631 305.529  < 2e-16 ***
## PARITY2        59.741     15.758   3.791 0.000151 ***
## PARITY3        20.638     18.069   1.142 0.253407
## PARITY4        -3.348     22.411  -0.149 0.881245
## PARITY5        -9.902     28.086  -0.353 0.724428
## PARITY6        -5.853     40.219  -0.146 0.884296
## PARITY7       -66.487     55.453  -1.199 0.230566
## PARITY8      -120.272     85.245  -1.411 0.158304
## PARITY9       -27.352    111.101  -0.246 0.805542
## PARITY10     -194.226    149.718  -1.297 0.194564
## PARITY11     -133.861    164.908  -0.812 0.416965
## PARITY12      -36.932    251.602  -0.147 0.883302
## PARITY13       19.068    275.575   0.069 0.944838
## PARITY14     -201.432    308.056  -0.654 0.513203
## PARITY15     -256.932    435.528  -0.590 0.555248
## PARITY16    -1234.932    435.528  -2.835 0.004585 **
## PARITY19     -822.932    615.837  -1.336 0.181488
## PARITY25      834.068    615.837   1.354 0.175651
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 615.7 on 9966 degrees of freedom
##   (16 observations deleted due to missingness)
## Multiple R-squared:  0.00394,    Adjusted R-squared:  0.002241
## F-statistic: 2.319 on 17 and 9966 DF,  p-value: 0.001583
```

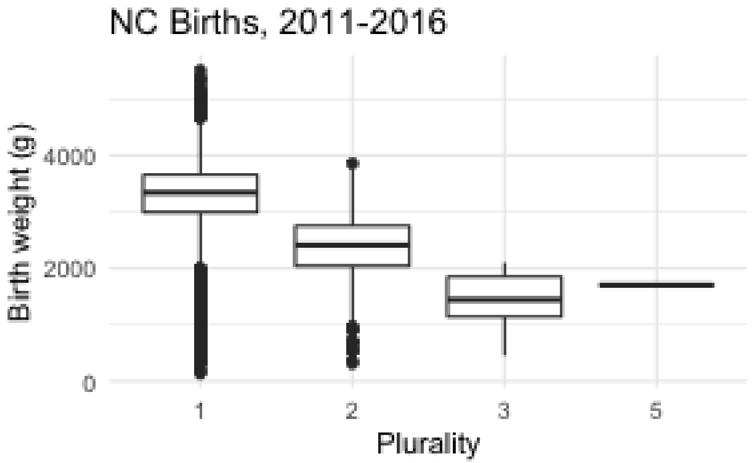


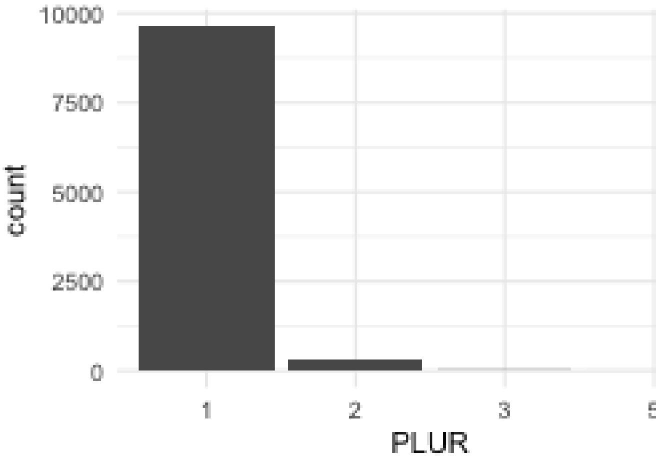


```
##
## Call:
## lm(formula = BWTG ~ PARITY_truncated, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3080.37  -313.09    46.43   384.42  2260.42
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3247.932     10.632 305.482  < 2e-16 ***
## PARITY_truncated2    59.741     15.761   3.791 0.000151 ***
## PARITY_truncated3    20.638     18.072   1.142 0.253480
## PARITY_truncated4    -3.348     22.415  -0.149 0.881263
## PARITY_truncated5+  -29.567     21.568  -1.371 0.170456
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 615.8 on 9979 degrees of freedom
##   (16 observations deleted due to missingness)
## Multiple R-squared:  0.002333,   Adjusted R-squared:  0.001933
## F-statistic: 5.834 on 4 and 9979 DF,  p-value: 0.0001097
```

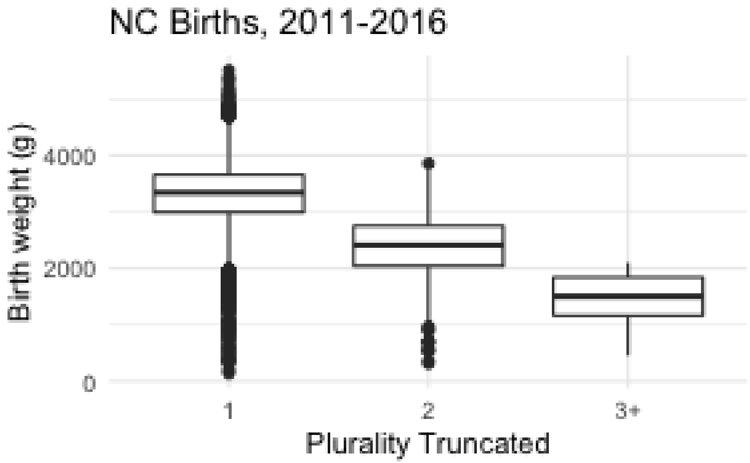
Independent of other variables, we see a negative relationship between parity and birth weight past the first child. The frequency of parity decreases in an exponential fashion. A second variable was created that truncates parities of at least five to improve interprability and prevent overfitting. The quantity of missing data is relatively small.

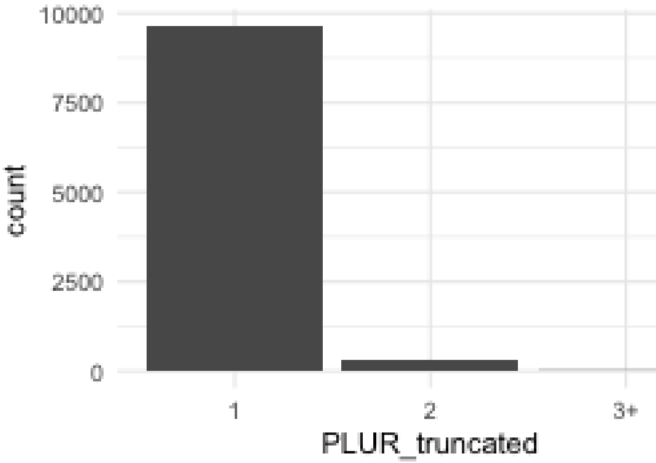
Plurality





```
##
## Call:
## lm(formula = BWTG ~ PLUR, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3161.43  -309.43    40.57   357.57  2205.57
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3299.430      5.988 550.969  < 2e-16 ***
## PLUR2        -959.607     32.490 -29.535  < 2e-16 ***
## PLUR3       -1852.597    169.836 -10.908  < 2e-16 ***
## PLUR5       -1598.430    587.994  -2.718  0.00657 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 588 on 9988 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.09053,    Adjusted R-squared:  0.09026
## F-statistic: 331.4 on 3 and 9988 DF,  p-value: < 2.2e-16
```

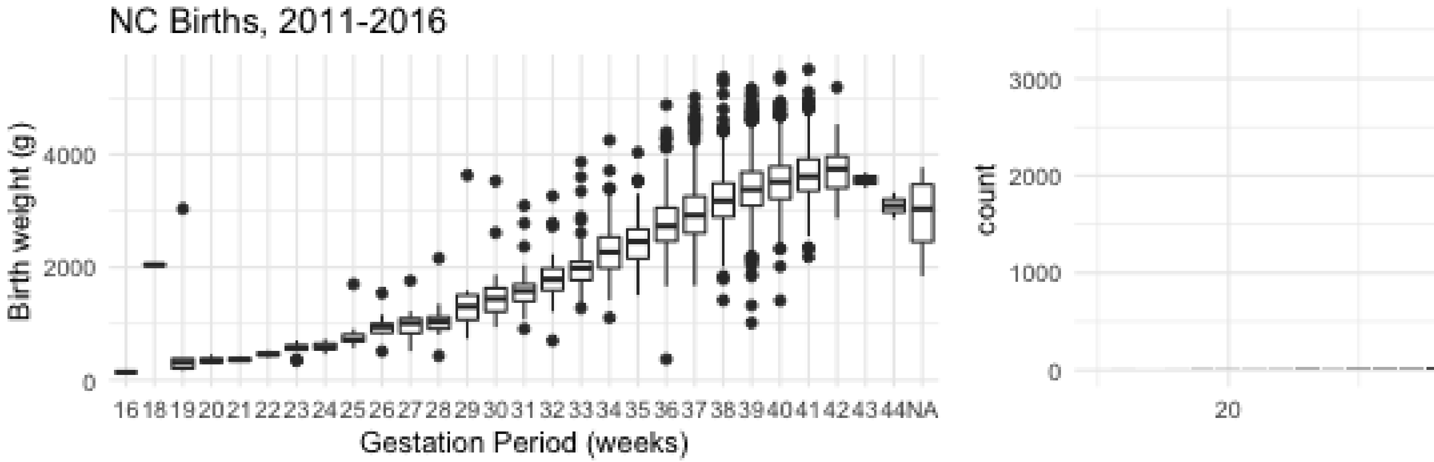




```
##
## Call:
## lm(formula = BWTG ~ PLUR_truncated, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3161.43  -309.43    40.57   357.57  2205.57
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3299.430      5.988  550.99   <2e-16 ***
## PLUR_truncated2   -959.607     32.489  -29.54   <2e-16 ***
## PLUR_truncated3+ -1833.046    163.175  -11.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 587.9 on 9989 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.09052,    Adjusted R-squared:  0.09033
## F-statistic: 497.1 on 2 and 9989 DF,  p-value: < 2.2e-16
```

We see a strong non linear negative relationship between plurality and birth weight. The frequency of pluralities above two is extremely small, and we again see a proportionally small amount of missing data. A second variable was created that truncates pluralities of at least three to improve interprability and prevent overfitting

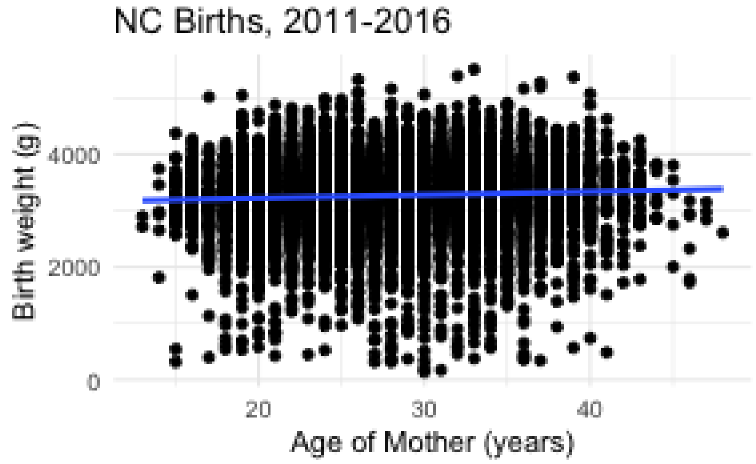
Gestation

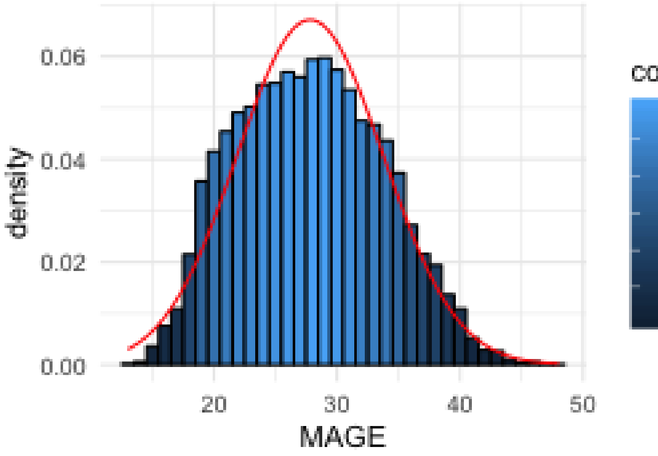


```
##
## Call:
## lm(formula = BWTG ~ GEST, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2418.6  -307.1   -26.7   281.9  3412.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3913.483     79.331  -49.33   <2e-16 ***
## GEST          186.168      2.054   90.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 456.5 on 9984 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.4514, Adjusted R-squared:  0.4513
## F-statistic:  8215 on 1 and 9984 DF,  p-value: < 2.2e-16
```

There appears to be a non linear positive relationship between gestation period and birth weight. The mean gestational period is approximately 38.5 weeks and the period with the highest median weight is 42 weeks. The frequency distribution is left skewed with the majority of babies having a gestational period between 38 and 40 weeks. There is some concern that more extreme gestational periods may lead to higher variance, and it should be noted that there is a chunk of data points with gestational periods of 17 to 21 weeks that have much higher than expected birth weights. There is an extreme outlier with gestational age of 83 weeks. Given that this data point was probably incorrectly recorded, we will exclude it from our analysis when building the model.

Age of Mother

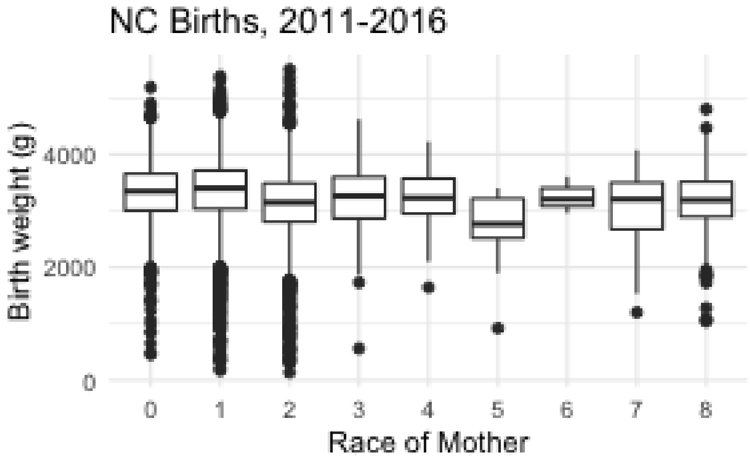




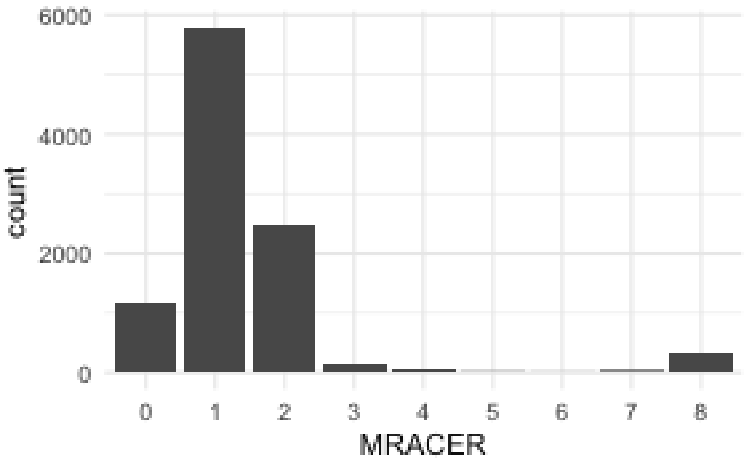
```
##
## Call:
## lm(formula = BWTG ~ MAGE, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3139.0  -311.9    51.5   386.3  2210.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3105.759     29.463 105.410  < 2e-16 ***
## MAGE           5.708      1.036   5.509 3.71e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 615.6 on 9989 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.003029,   Adjusted R-squared:  0.002929
## F-statistic: 30.35 on 1 and 9989 DF,  p-value: 3.706e-08
```

Mother's age seems to be fairly normally distributed with a mean of 27.8. There appears to be a positive relationship between the age of the mother and the birth weight. There is no evidence to suggest that the birth weight variance is not constant across the mother's age.

Race of Mother



```
## # A tibble: 9 x 3
##   MRACER mweight freq_percent
##   <fct>    <dbl>        <dbl>
## 1 0        3313.     0.118
## 2 1        3340.     0.579
## 3 2        3083.     0.246
## 4 3        3224.     0.0145
## 5 4        3220.     0.005
## 6 5        2622.     0.0008
## 7 6        3254.     0.000300
## 8 7        3045.     0.0028
## 9 8        3176.     0.0329
```



```
##
## Call:
## lm(formula = BWTG ~ MRACER, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3170.0  -307.0    51.2   375.1  2421.6
##
```

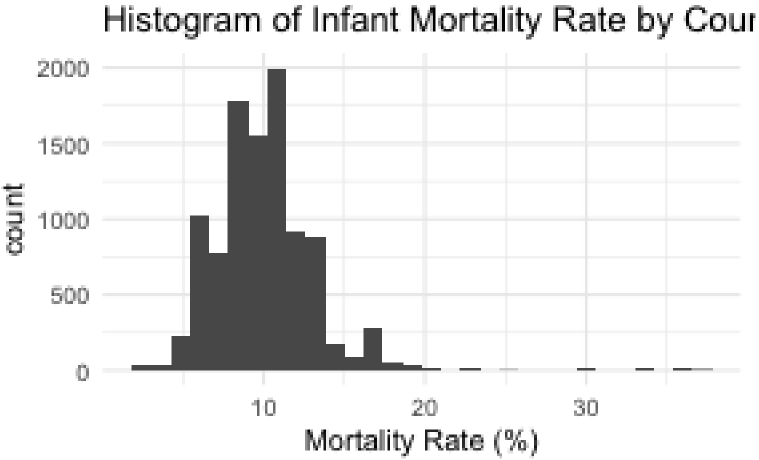
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3313.26      17.62 188.047  < 2e-16 ***
## MRACER1        26.73      19.34   1.382 0.166914
## MRACER2      -229.87      21.44 -10.720  < 2e-16 ***
## MRACER3       -88.81      53.36  -1.664 0.096088 .
## MRACER4       -92.96      87.57  -1.062 0.288460
## MRACER5      -691.51     215.16  -3.214 0.001314 **
## MRACER6       -58.92     350.62  -0.168 0.866542
## MRACER7      -268.44     115.97  -2.315 0.020648 *
## MRACER8      -137.31      37.80  -3.633 0.000282 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 606.5 on 9983 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.03269,    Adjusted R-squared:  0.03191
## F-statistic: 42.17 on 8 and 9983 DF,  p-value: < 2.2e-16
```

0 - non-white„ 1 = white, 2 black, 3 indian, 8 other asian

There are significant differences between the average birth weights of mother's of different races. We see that mother's that self identified as white have the largest mean baby weight at 3.33 kg, while black mother's have the lowest mean baby weight at only 3.07 kg. 58 percent of mother's identify as white, 24 percent identify as black, 12 percent identify as non-white, and 3 percent identify as other asian.

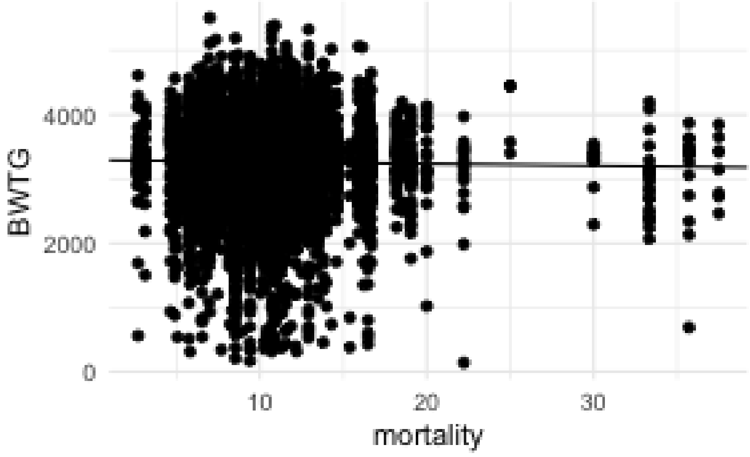
County / Socioeconomic Status

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.703   8.532   9.459  10.049  11.561  37.500
```



```
##
## Call:
## lm(formula = BWTG ~ mortality, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3096.40  -315.83    48.09   387.44  2231.64
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3293.701     19.633 167.761   <2e-16 ***
## mortality     -2.905      1.854  -1.567    0.117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 616.5 on 9941 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.0002469,  Adjusted R-squared:  0.0001463
## F-statistic: 2.455 on 1 and 9941 DF,  p-value: 0.1172
```

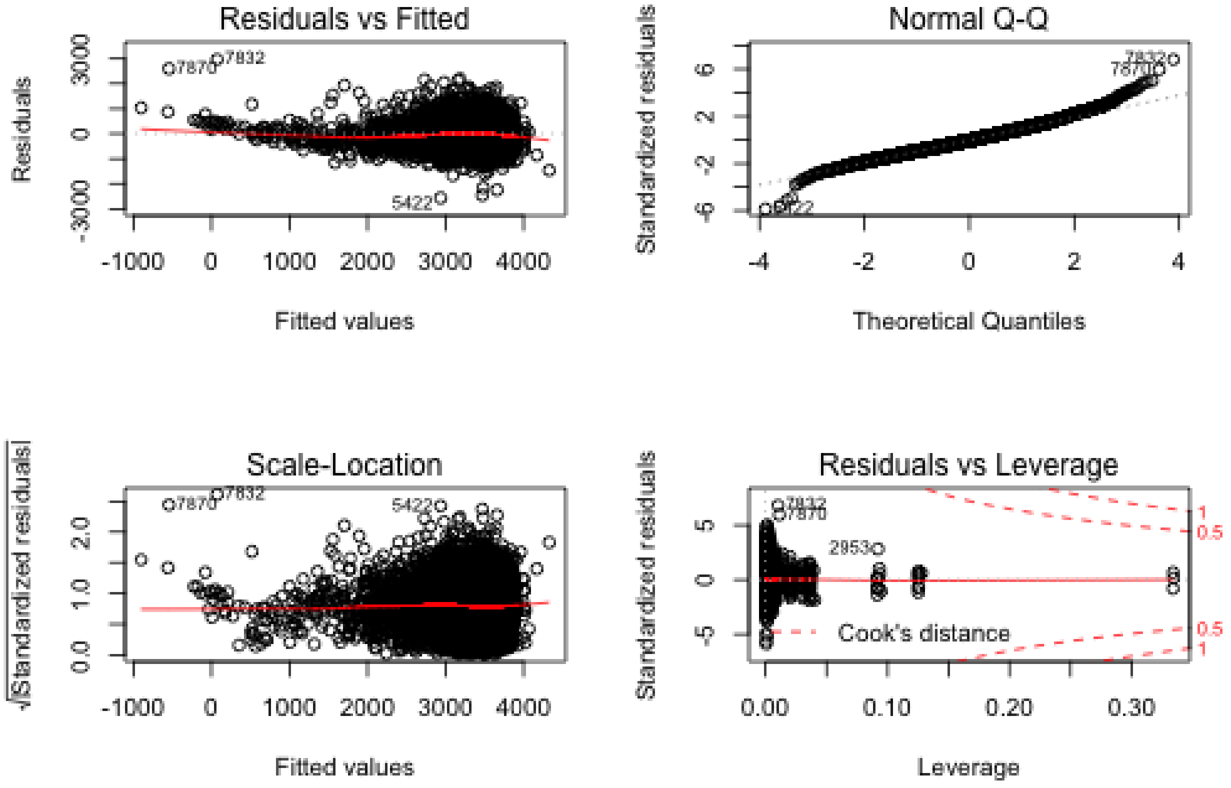


We chose to use infant mortality rate of birth county as a proxy for socioeconomic status, calculated as number of deaths before the age of 1 divided by total number of births in a county. The median county in the data had a infant mortality rate of 0.7%, with the range of infant mortality rates in our dataset ranging from 0.12% to 1.76%. Infant mortality rate of birth county and birth weight appear to have a weak negative linear relationship, and in isolation, a 1 percentage point increase in infant mortality rate is associated with a 157g decrease in expected birth weight. ## Build Model

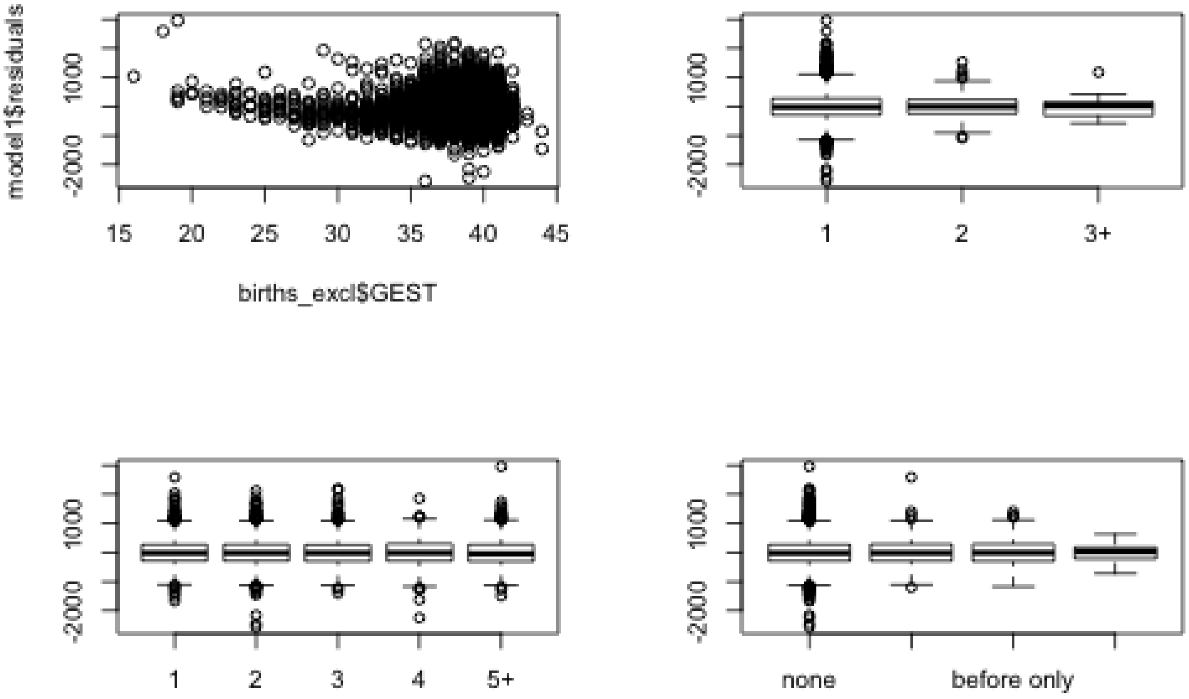
```
births_excl = na.omit(births)
births_excl = births_excl[which(births_excl$GEST < 80), ]
births_excl = births_excl %>%
  mutate(GEST2 = GEST^2, GEST3 = GEST ^ 3, GEST4 = GEST^4)
model1 = lm(data = births_excl, BWTG ~ GEST + PARITY_truncated + PLUR_truncated + smoking_type + MAGE +
summary(model1)
```

```
##
## Call:
## lm(formula = BWTG ~ GEST + PARITY_truncated + PLUR_truncated +
##     smoking_type + MAGE + MRACER + mortality, data = births_excl,
##     na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2565.06  -293.07   -22.57   267.69  2963.26
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -3674.1210    87.8290 -41.833  < 2e-16 ***
## GEST                            174.9507     2.1122  82.829  < 2e-16 ***
## PARITY_truncated2                86.5300    11.4807   7.537 5.23e-14 ***
## PARITY_truncated3               101.3744    13.4307   7.548 4.81e-14 ***
## PARITY_truncated4                82.0486    16.7811   4.889 1.03e-06 ***
## PARITY_truncated5+              116.4825    16.7717   6.945 4.02e-12 ***
## PLUR_truncated2                -366.8503    25.4843 -14.395  < 2e-16 ***
## PLUR_truncated3+               -330.6853   133.2559  -2.482   0.0131 *
## smoking_typebefore and during  -223.2632    15.1174 -14.769  < 2e-16 ***
## smoking_typebefore only           7.7057    23.9234   0.322   0.7474
## smoking_typeduring only        -149.9606    76.3179  -1.965   0.0494 *
## MAGE                              5.1549     0.8298   6.213 5.42e-10 ***
## MRACER1                          63.3735    14.3478   4.417 1.01e-05 ***
## MRACER2                        -117.0007    15.6532  -7.475 8.40e-14 ***
## MRACER3                          15.9737    38.9232   0.410   0.6815
## MRACER4                        -114.9173    63.2444  -1.817   0.0692 .
## MRACER5                        -368.9921   155.2759  -2.376   0.0175 *
## MRACER6                        -106.0637   252.5700  -0.420   0.6745
## MRACER7                         -44.4303    83.6917  -0.531   0.5955
## MRACER8                        -132.7009    27.4838  -4.828 1.40e-06 ***
## mortality                         1.9595     1.3237   1.480   0.1388
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 436.8 on 9879 degrees of freedom
## Multiple R-squared:  0.4974, Adjusted R-squared:  0.4963
## F-statistic: 488.8 on 20 and 9879 DF,  p-value: < 2.2e-16
```

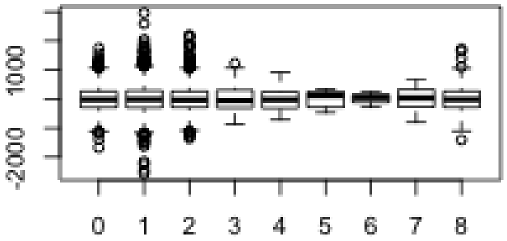
```
par(mfrow = c(2,2))
plot(model1)
```

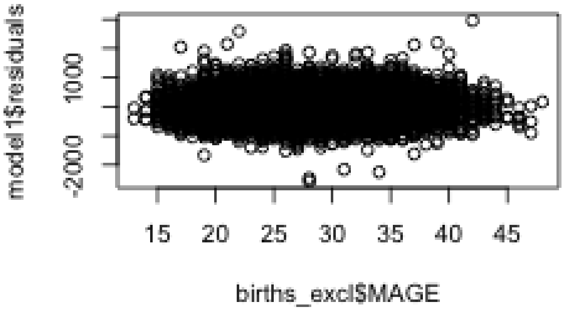


```r
# plot(model1$fitted.values, model1$residuals)
plot(births_excl$GEST, model1$residuals)
plot(births_excl$PLUR_truncated, model1$residuals)
plot(births_excl$PARITY_truncated, model1$residuals)
plot(births_excl$smoking_type, model1$residuals)
```



```
plot(births_excl$MRACER, model1$residuals)
plot(births_excl$MAGE, model1$residuals)
```





The residuals vs fitted values and residuals vs gestational period plot slope downwards, indicating that there is a departure from linearity. More precisely, the linear model underpredicts when gestational period is below ~30 and overpredicts when gestational period is above ~30. A transformation may be helpful. The model may improve if a square term is added. There is a particularly high residual (in terms of absolute value) around 80 weeks of gestation, which is likely an outlier that has no reason to be there, as no humans can possibly gestate for 80 weeks (~1.54 years).

The residual graph for Plurality (truncated) has decreasing residuals (in terms of absolute value) as plurality increases. This makes sense, as birth weight should get smaller (and as a result range of birth weights should get tighter, leading to smaller absolute value residuals) as more babies share a womb and share nutrients – More sharing will biologically cause them to come out smaller.

The residual graph for Parity (truncated) has pretty random residuals that are all around the same size for each group.

The residual graph for Smoking has higher residuals for no smoking than for smoking of any kind. This makes sense, as birth weight could biologically get smaller in the presence of smoking, as smoking can be damaging to the fetus and be detrimental to its growth and weight. This would lead to the range of birth weights of smoking mothers getting tighter, leading to smaller absolute value residuals.

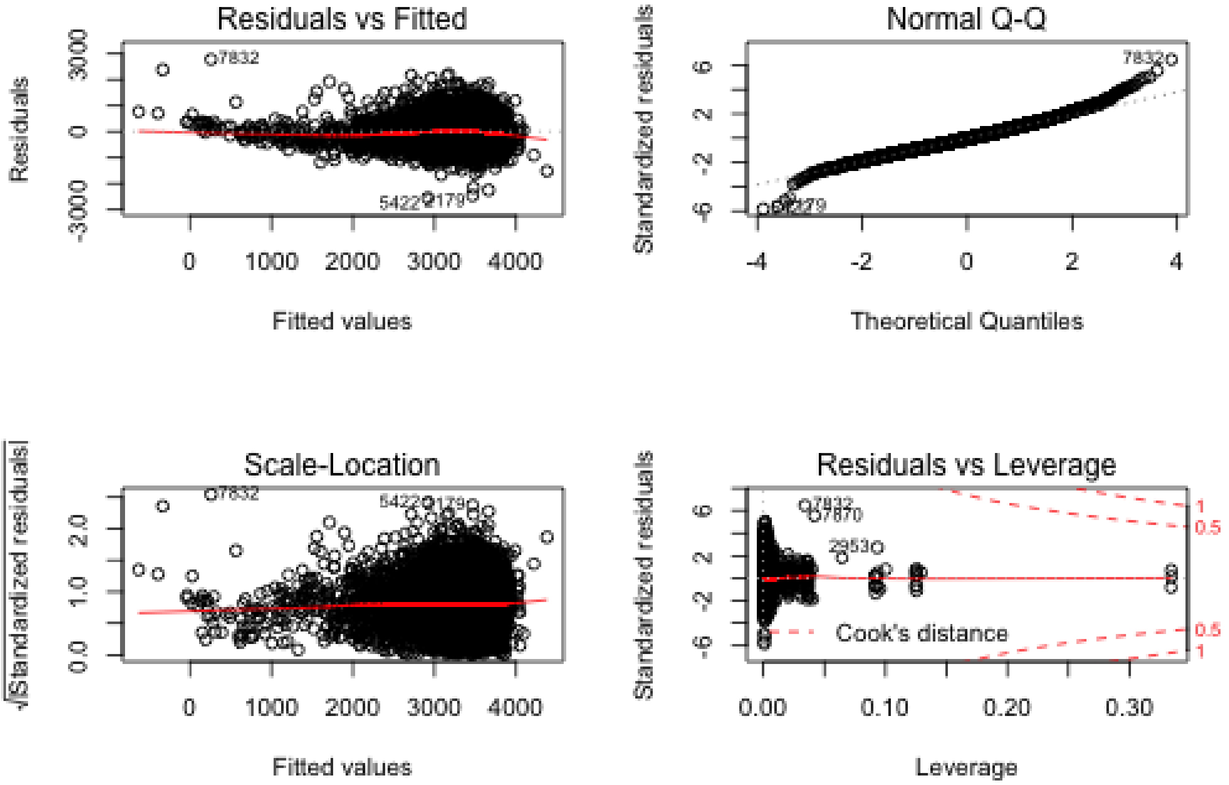
The residual graph for Mother's race indicates that residuals are lower for for races 3, 4, 5, 6, and 7 and higher for the other races. This could be something to explore.

The residual graph for Mother's age is fairly random, with residuals getting a bit smaller near the beginning and end (<20 years old and >45 years old).

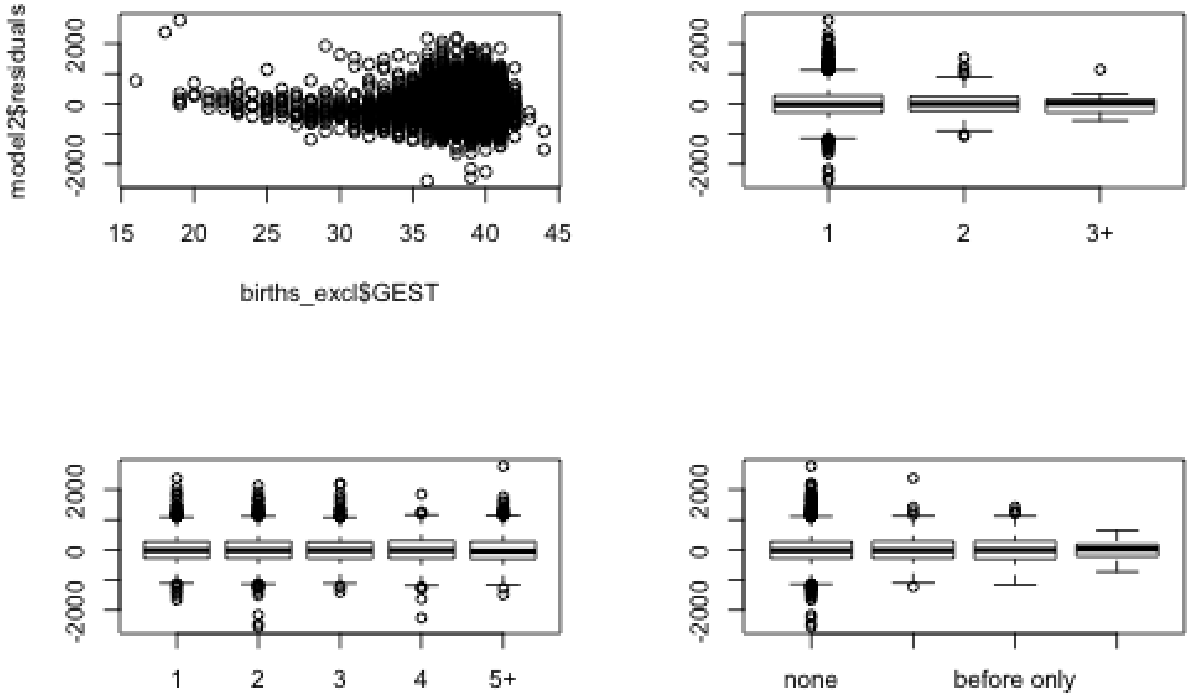
```
model2 = lm(data = births_excl, BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated + smoking_type
summary(model2)
```

```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated +
##     smoking_type + MAGE + MRACER + mortality, data = births_excl,
##     na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2550.41  -292.32   -23.35   269.28  2775.25
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2703.3592   360.0375  -7.509 6.50e-14 ***
## GEST                            118.0489    20.5753   5.737 9.90e-09 ***
## GEST2                             0.8187     0.2945   2.780  0.00544 **
## PARITY_truncated2                87.9431    11.4881   7.655 2.11e-14 ***
## PARITY_truncated3               102.7765    13.4356   7.650 2.21e-14 ***
## PARITY_truncated4                83.5634    16.7842   4.979 6.51e-07 ***
## PARITY_truncated5+              118.1519    16.7767   7.043 2.01e-12 ***
## PLUR_truncated2                -359.5892    25.6091 -14.041  < 2e-16 ***
## PLUR_truncated3+               -348.4866   133.3643  -2.613  0.00899 **
## smoking_typebefore and during  -222.5133    15.1146 -14.722  < 2e-16 ***
## smoking_typebefore only           7.0272    23.9165   0.294  0.76890
## smoking_typeduring only        -148.6503    76.2933  -1.948  0.05139 .
## MAGE                              5.2004     0.8296   6.268 3.80e-10 ***
## MRACER1                          62.8178    14.3443   4.379 1.20e-05 ***
## MRACER2                        -117.3107    15.6482  -7.497 7.10e-14 ***
## MRACER3                          15.6184    38.9102   0.401  0.68814
## MRACER4                        -114.5448    63.2231  -1.812  0.07005 .
## MRACER5                        -371.7825   155.2263  -2.395  0.01663 *
## MRACER6                        -106.3402   252.4841  -0.421  0.67364
## MRACER7                         -40.2203    83.6769  -0.481  0.63077
## MRACER8                        -132.2184    27.4750  -4.812 1.51e-06 ***
## mortality                         1.9449     1.3232   1.470  0.14164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 436.6 on 9878 degrees of freedom
## Multiple R-squared:  0.4978, Adjusted R-squared:  0.4967
## F-statistic: 466.2 on 21 and 9878 DF,  p-value: < 2.2e-16
```

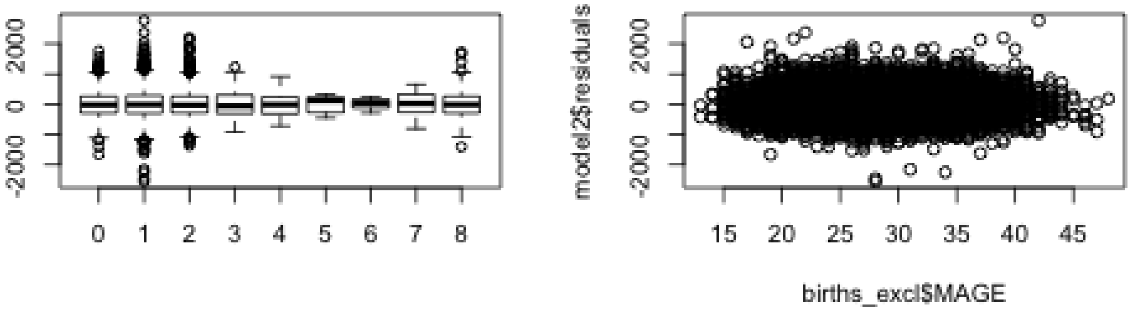
```
par(mfrow = c(2,2))
plot(model2)
```



```r
# plot(model2$fitted.values, model2$residuals)
plot(births_excl$GEST, model2$residuals)
plot(births_excl$PLUR_truncated, model2$residuals)
plot(births_excl$PARITY_truncated, model2$residuals)
plot(births_excl$smoking_type, model2$residuals)
```



```
plot(births_excl$MRACER, model2$residuals)
plot(births_excl$MAGE, model2$residuals)
```

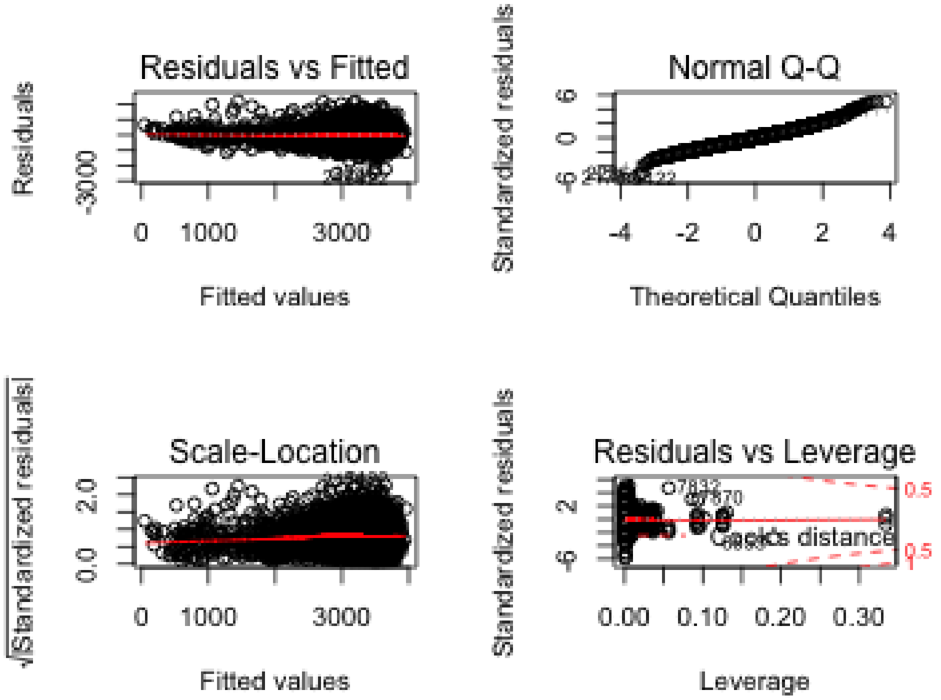


The new model still displays the original downwards trend in the residual vs gestational period graph. Perhaps another transformation on GEST would be helpful – a cubic term can be added. The other residuals plots also retain their trends from model 1.

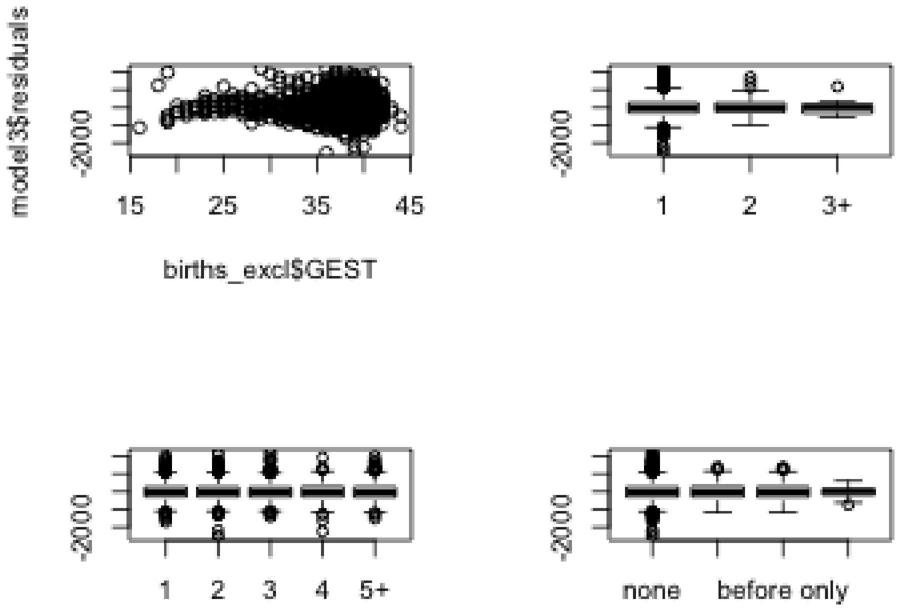
```
model3 = lm(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated + PLUR_truncated + smoki
summary(model3)
```

```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2548.79  -288.43   -24.73   266.11  2185.79
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.644e+04  1.589e+03  10.344  < 2e-16 ***
## GEST                         -1.730e+03  1.509e+02 -11.465  < 2e-16 ***
## GEST2                         5.841e+01  4.668e+00  12.512  < 2e-16 ***
## GEST3                        -5.838e-01  4.723e-02 -12.361  < 2e-16 ***
## PARITY_truncated2             8.254e+01  1.141e+01   7.234 5.04e-13 ***
## PARITY_truncated3             9.695e+01  1.334e+01   7.267 3.96e-13 ***
## PARITY_truncated4             7.430e+01  1.667e+01   4.456 8.43e-06 ***
## PARITY_truncated5+              1.104e+02  1.666e+01   6.628 3.57e-11 ***
## PLUR_truncated2                -3.261e+02  2.556e+01 -12.760  < 2e-16 ***
## PLUR_truncated3+               -1.902e+02  1.330e+02  -1.431   0.1526    
## smoking_typebefore and during -2.236e+02  1.500e+01 -14.908  < 2e-16 ***
## smoking_typebefore only         9.569e+00  2.374e+01   0.403   0.6868    
## smoking_typeduring only        -1.525e+02  7.571e+01  -2.014   0.0440 *  
## MAGE                            5.063e+00  8.234e-01   6.148 8.13e-10 ***
## MRACER1                         6.266e+01  1.424e+01   4.402 1.08e-05 ***
## MRACER2                        -1.168e+02  1.553e+01  -7.522 5.88e-14 ***
## MRACER3                         2.179e+01  3.862e+01   0.564   0.5726    
## MRACER4                        -1.153e+02  6.274e+01  -1.838   0.0660 .  
## MRACER5                        -3.492e+02  1.541e+02  -2.266   0.0234 *  
## MRACER6                        -9.661e+01  2.506e+02  -0.386   0.6998    
## MRACER7                        -3.378e+01  8.304e+01  -0.407   0.6841    
## MRACER8                        -1.348e+02  2.727e+01  -4.945 7.74e-07 ***
## mortality                       1.521e+00  1.314e+00   1.158   0.2468    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 433.3 on 9877 degrees of freedom
## Multiple R-squared:  0.5054, Adjusted R-squared:  0.5043
## F-statistic: 458.8 on 22 and 9877 DF,  p-value: < 2.2e-16
```

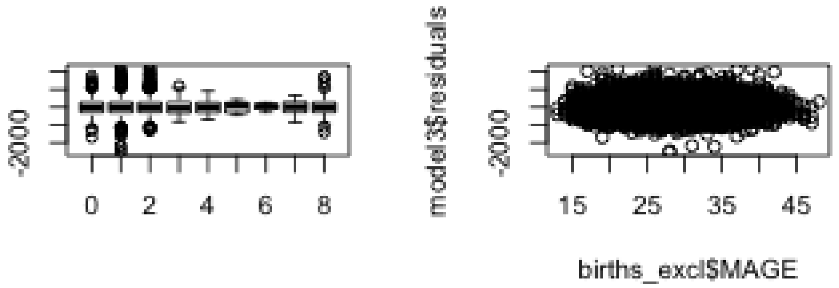
```r
par(mfrow = c(2,2))
plot(model3)
```



```r
# plot(model3$fitted.values, model3$residuals)
plot(births_excl$GEST, model3$residuals)
plot(births_excl$PLUR_truncated, model3$residuals)
plot(births_excl$PARITY_truncated, model3$residuals)
plot(births_excl$smoking_type, model3$residuals)
```



```
plot(births_excl$MRACER, model3$residuals)
plot(births_excl$MAGE, model3$residuals)
```

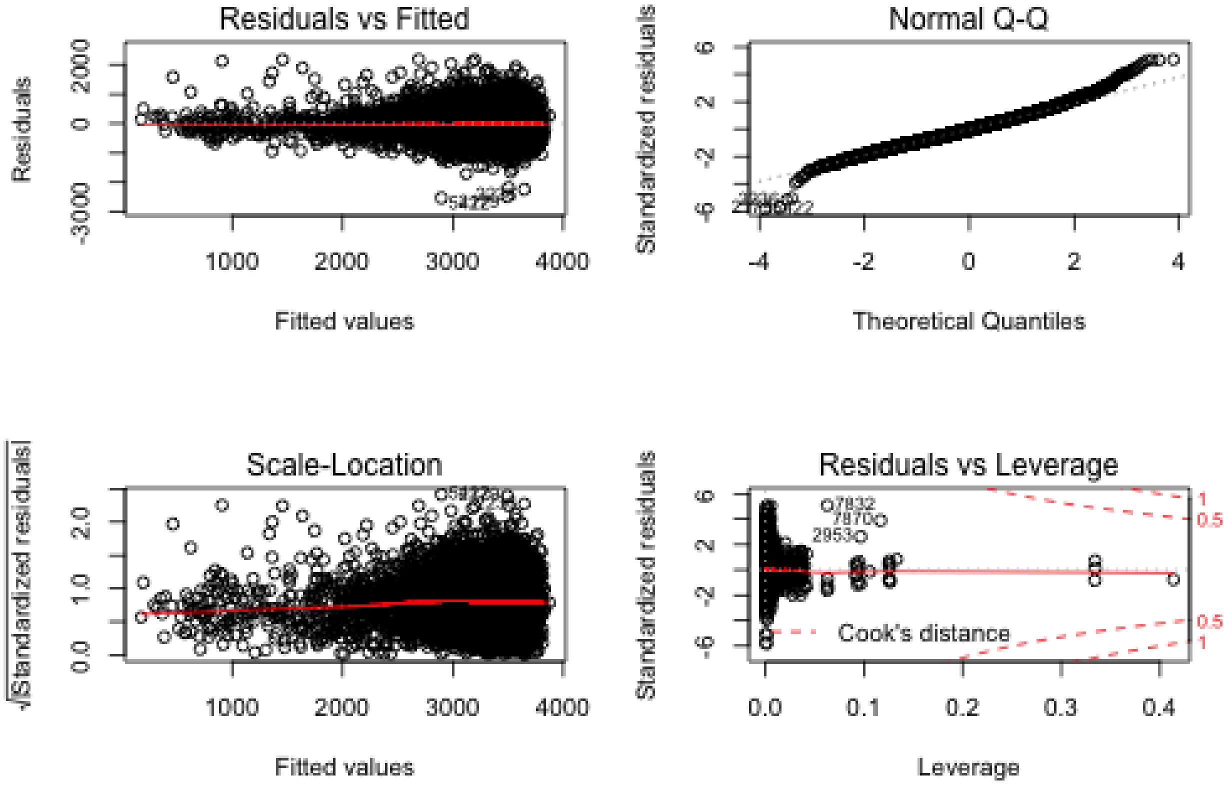


The residual vs gestational period shows a much more random pattern than before. It is worth investigating if adding a quartic term would help. The other residuals plots also retain their trends from model 1.

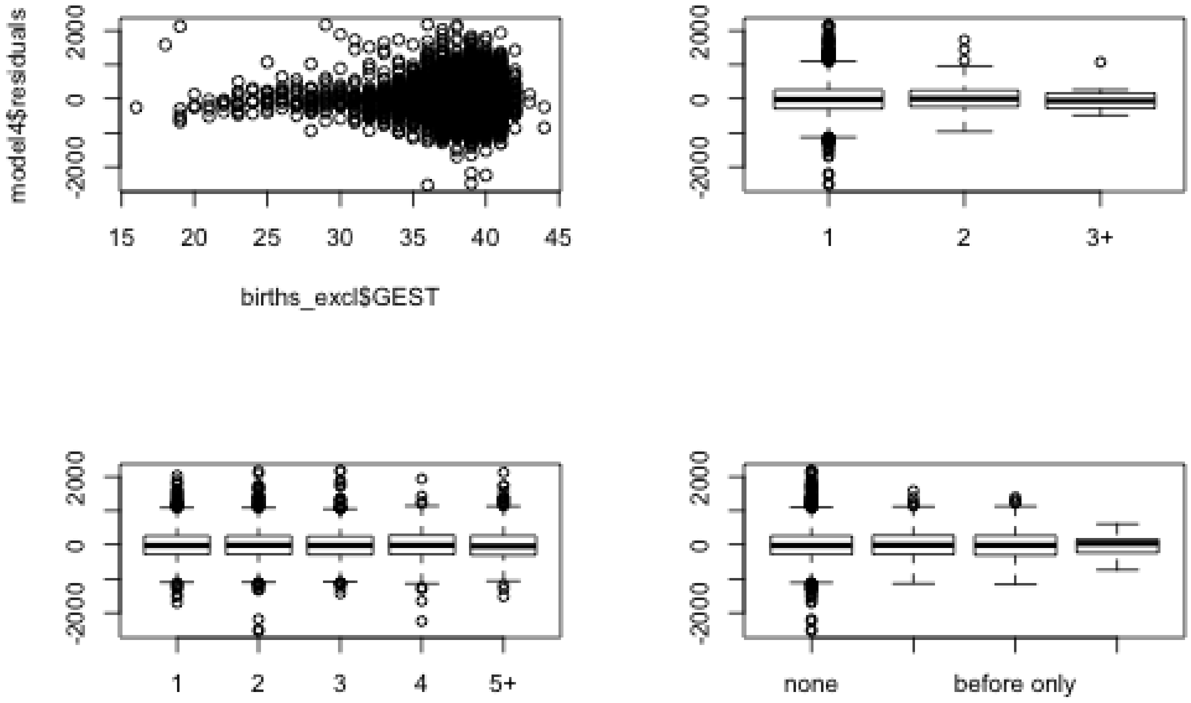
```
model4 = lm(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_truncated
summary(model4)
```

```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2527.63  -288.51   -24.22   265.01  2182.41
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.034e+04  6.399e+03  -1.616 0.106147
## GEST                          1.880e+03  8.490e+02   2.214 0.026845 *
## GEST2                        -1.190e+02  4.133e+01  -2.880 0.003991 **
## GEST3                         3.201e+00  8.773e-01   3.649 0.000265 ***
## GEST4                        -2.968e-02  6.869e-03  -4.320 1.57e-05 ***
## PARITY_truncated2             8.015e+01  1.141e+01   7.023 2.32e-12 ***
## PARITY_truncated3             9.384e+01  1.335e+01   7.029 2.22e-12 ***
## PARITY_truncated4             7.233e+01  1.666e+01   4.340 1.44e-05 ***
## PARITY_truncated5+            1.107e+02  1.665e+01   6.648 3.13e-11 ***
## PLUR_truncated2              -3.138e+02  2.570e+01 -12.210  < 2e-16 ***
## PLUR_truncated3+             -2.079e+02  1.329e+02  -1.564 0.117848
## smoking_typebefore and during -2.222e+02  1.499e+01 -14.821  < 2e-16 ***
## smoking_typebefore only       8.732e+00  2.372e+01   0.368 0.712745
## smoking_typeduring only      -1.576e+02  7.566e+01  -2.083 0.037307 *
## MAGE                          5.060e+00  8.227e-01   6.151 8.01e-10 ***
## MRACER1                       6.203e+01  1.422e+01   4.361 1.31e-05 ***
## MRACER2                      -1.178e+02  1.552e+01  -7.593 3.40e-14 ***
## MRACER3                       2.332e+01  3.858e+01   0.604 0.545546
## MRACER4                      -1.173e+02  6.269e+01  -1.871 0.061384 .
## MRACER5                      -3.621e+02  1.539e+02  -2.352 0.018695 *
## MRACER6                      -9.185e+01  2.503e+02  -0.367 0.713712
## MRACER7                      -3.339e+01  8.297e+01  -0.402 0.687405
## MRACER8                      -1.359e+02  2.724e+01  -4.989 6.18e-07 ***
## mortality                     1.550e+00  1.312e+00   1.181 0.237630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432.9 on 9876 degrees of freedom
## Multiple R-squared:  0.5063, Adjusted R-squared:  0.5052
## F-statistic: 440.4 on 23 and 9876 DF,  p-value: < 2.2e-16
```

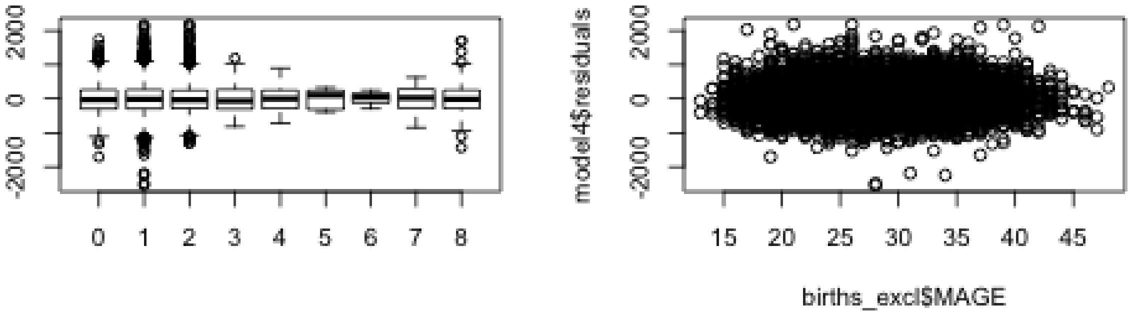
```
par(mfrow = c(2,2))
plot(model4)
```



```r
# plot(model4$fitted.values, model4$residuals)
plot(births_excl$GEST, model4$residuals)
plot(births_excl$PLUR_truncated, model4$residuals)
plot(births_excl$PARITY_truncated, model4$residuals)
plot(births_excl$smoking_type, model4$residuals)
```



```r
plot(births_excl$MRACER, model4$residuals)
plot(births_excl$MAGE, model4$residuals)
```



The addition of the quartic term does not seem to help. The residual vs gestational period graph shows that residuals increase in absolute value as gestational period increases from 20 to 40 weeks. These residuals are much less random than that of model 3.

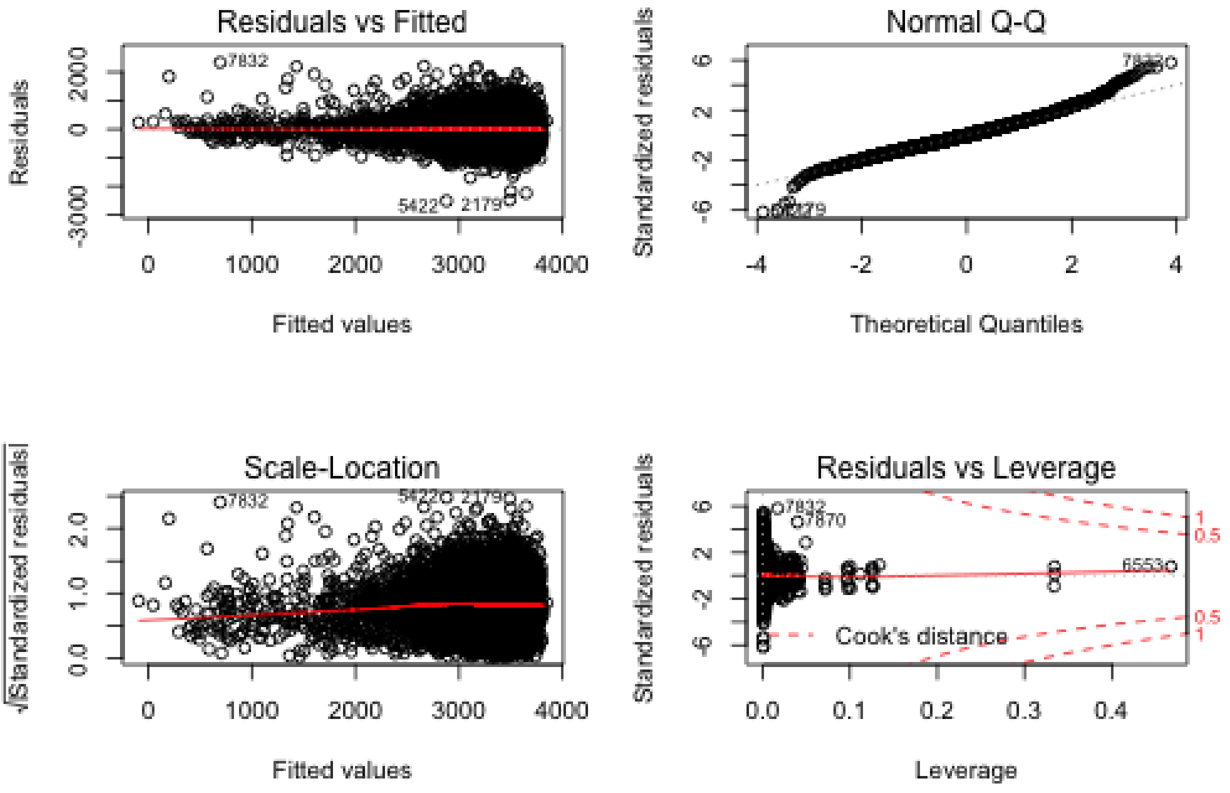
Model 3 looks like the best, but perhaps we can use robust regression to improve upon this the massive residual of the outlier point near 80 weeks of gestational age.

Robust on Model 4

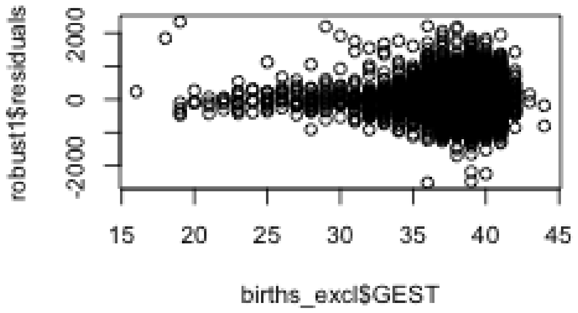
```
robust1 <- rlm(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_trunca
summary(robust1)
```

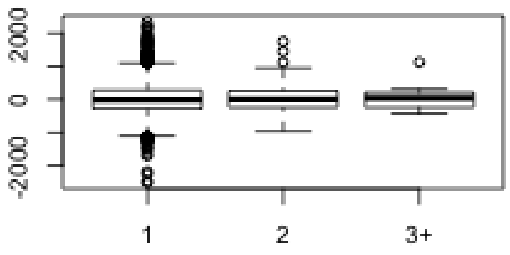
```
##
## Call: rlm(formula = BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2508.86  -276.15   -13.13   274.07  2343.00
##
## Coefficients:
##                                 Value       Std. Error  t value
## (Intercept)                     -18663.6717   6227.3839     -2.9970
## GEST                              2931.0654    826.2896      3.5473
## GEST2                             -168.3403     40.2223     -4.1852
## GEST3                               4.2167      0.8538      4.9386
## GEST4                              -0.0374      0.0067     -5.5964
## PARITY_truncated2                  82.9965     11.1065      7.4728
## PARITY_truncated3                  90.4003     12.9916      6.9583
## PARITY_truncated4                  81.5808     16.2183      5.0302
## PARITY_truncated5+                102.5824     16.2000      6.3323
## PLUR_truncated2                  -289.9793     25.0071    -11.5959
## PLUR_truncated3+                 -223.5662    129.3503     -1.7284
## smoking_typebefore and during    -210.7046     14.5886    -14.4431
## smoking_typebefore only             4.1363     23.0796      0.1792
## smoking_typeduring only          -138.8395     73.6279     -1.8857
## MAGE                                4.8441      0.8006      6.0504
## MRACER1                            63.8905     13.8422      4.6156
## MRACER2                          -123.2072     15.1014     -8.1587
## MRACER3                            11.3276     37.5507      0.3017
## MRACER4                          -110.9609     61.0080     -1.8188
## MRACER5                          -346.3442    149.8228     -2.3117
## MRACER6                           -76.6610    243.6355     -0.3147
## MRACER7                            -4.0102     80.7449     -0.0497
## MRACER8                          -135.8743     26.5137     -5.1247
## mortality                           1.5249      1.2773      1.1939
##
## Residual standard error: 407.7 on 9876 degrees of freedom
```

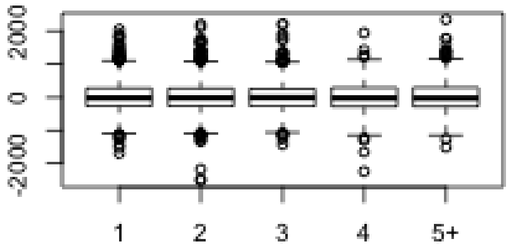
```
par(mfrow = c(2,2))
plot(robust1)
```

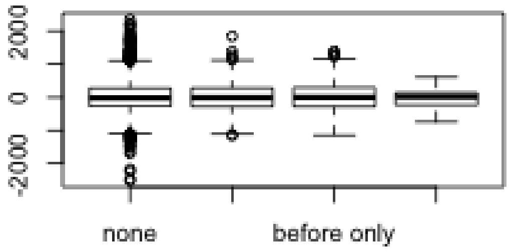


```r
# plot(robust1$fitted.values, robust1$residuals)
plot(births_excl$GEST, robust1$residuals)
plot(births_excl$PLUR_truncated, robust1$residuals)
plot(births_excl$PARITY_truncated, robust1$residuals)
plot(births_excl$smoking_type, robust1$residuals)
```





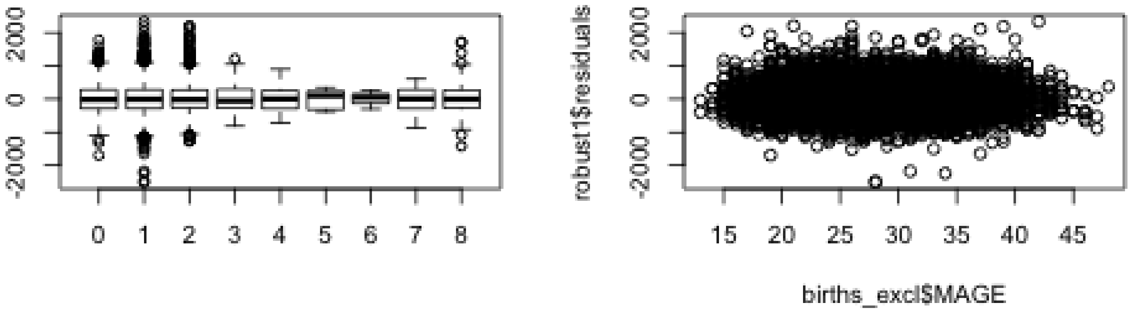




```r
plot(births_excl$MRACER, robust1$residuals)
plot(births_excl$MAGE, robust1$residuals)

#Check weights
robust1_weights = data.frame(bwt = births_excl$BWTG, gest = births_excl$GEST,
 resid=robust1$resid, weight=robust1$w)
robust1_weights[order(robust1$w)[c(1:5, (length(robust1$w)-5):length(robust1$w))],]
```

```
##       bwt gest       resid    weight
## 5422  370   36 -2508.86317 0.2185844
## 2179 1013   39 -2474.79111 0.2215947
## 7832 3036   19  2343.00368 0.2340747
## 2236 1411   40 -2239.57237 0.2448706
## 8407 4876   36  2213.14818 0.2477999
## 9893 2778   37  -411.69929 1.0000000
## 9894 3848   39   321.81425 1.0000000
## 9895 3427   40  -182.67294 1.0000000
## 9897 3147   38   -70.65164 1.0000000
## 9899 3656   40   323.16272 1.0000000
## 9900 3428   39    26.12302 1.0000000
```



Note: change below to indicate taking out outlier

Checking the weights, the outlier point at gest = 83 with the residual of 57619 has indeed been weighted down (with a weight of 0.0093). The weights of four other points with high residuals are also weighted down.

Looking at the residual plot for gestational period, the residuals look mostly random (ignoring the outlier point at gest = 83).

Cross Validation

```
births_cv<-births_excl[sample(nrow(births_excl)),]
folds<-cut(seq(1,nrow(births_cv)),breaks=10,labels=FALSE)
test_list<-list()
train_list<-list()
for(i in 1:10){
  test_indices<-which(folds==i,arr.ind=TRUE)
  births_test<-births_cv[test_indices,]
  test_list[[i]]<-births_test
  births_train<-births_cv[-test_indices,]
  train_list[[i]]<-births_train
}
```

```r
#Train and test model1
model1_test_mse<-list()
for(i in 1:10){
  model1_train<-lm(data=train_list[[i]],BWTG~GEST+PARITY_truncated+PLUR_truncated+smoking_type+MAGE+MRAC
  model1_test<-predict(model1_train,train_list[[i]])
  model1_test_mse[[i]]<-mean((train_list[[i]]$BWTG-model1_test)^2)
}
test_mse<-list(model1_test_mse)

#Train and test model2
model2_test_mse<-list()
for(i in 1:10){
  model2_train<-lm(data=train_list[[i]],BWTG~GEST+GEST2+PARITY_truncated+PLUR_truncated+smoking_type+MAG
  model2_test<-predict(model2_train,train_list[[i]])
  model2_test_mse[[i]]<-mean((train_list[[i]]$BWTG-model2_test)^2)
}
test_mse<-append(test_mse,list(model2_test_mse))

#Train and test model3
model3_test_mse<-list()
for(i in 1:10){
  model3_train<-lm(data=train_list[[i]],BWTG~GEST+GEST2+GEST3+PARITY_truncated+PLUR_truncated+smoking_ty
  model3_test<-predict(model3_train,train_list[[i]])
  model3_test_mse[[i]]<-mean((train_list[[i]]$BWTG-model3_test)^2)
}
test_mse<-append(test_mse,list(model3_test_mse))

#Train and test model4
model4_test_mse<-list()
for(i in 1:10){
  model4_train<-lm(data=train_list[[i]],BWTG~GEST+GEST2+GEST3+GEST4+PARITY_truncated+PLUR_truncated+smok
  model4_test<-predict(model4_train,train_list[[i]])
  model4_test_mse[[i]]<-mean((train_list[[i]]$BWTG-model4_test)^2)
}
test_mse<-append(test_mse,list(model4_test_mse))

Train and test robust1
 robust1_test_mse<-list()
# for(i in 1:10){
#   robust1_train<-rlm(data=train_list[[i]],BWTG~GEST+GEST2+GEST3+PARITY_truncated+PLUR_truncated+smoki
#   robust1_test<-predict(robust1_train,train_list[[i]])
#   robust1_test_mse[[i]]<-mean((train_list[[i]]$BWTG-robust1_test)^2)
# }
 test_mse<-append(test_mse,list(robust1_test_mse))

#Results
results_cv<-matrix(c(lapply(test_mse,mean)),ncol=5)
colnames(results_cv)<-c('model1','model2','model3','model4','robust1')
rownames(results_cv)<-c('Average MSE')
results<-as.table(results_cv)
results
```