440 Case Study I

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

9/3/2019

Set Up, Load, Clean Data

```r
## load packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(MASS)
```

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

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

```r
## 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
```

Exploratory Data Analysis

Birthweight

```
summary(births$BWTG)
```

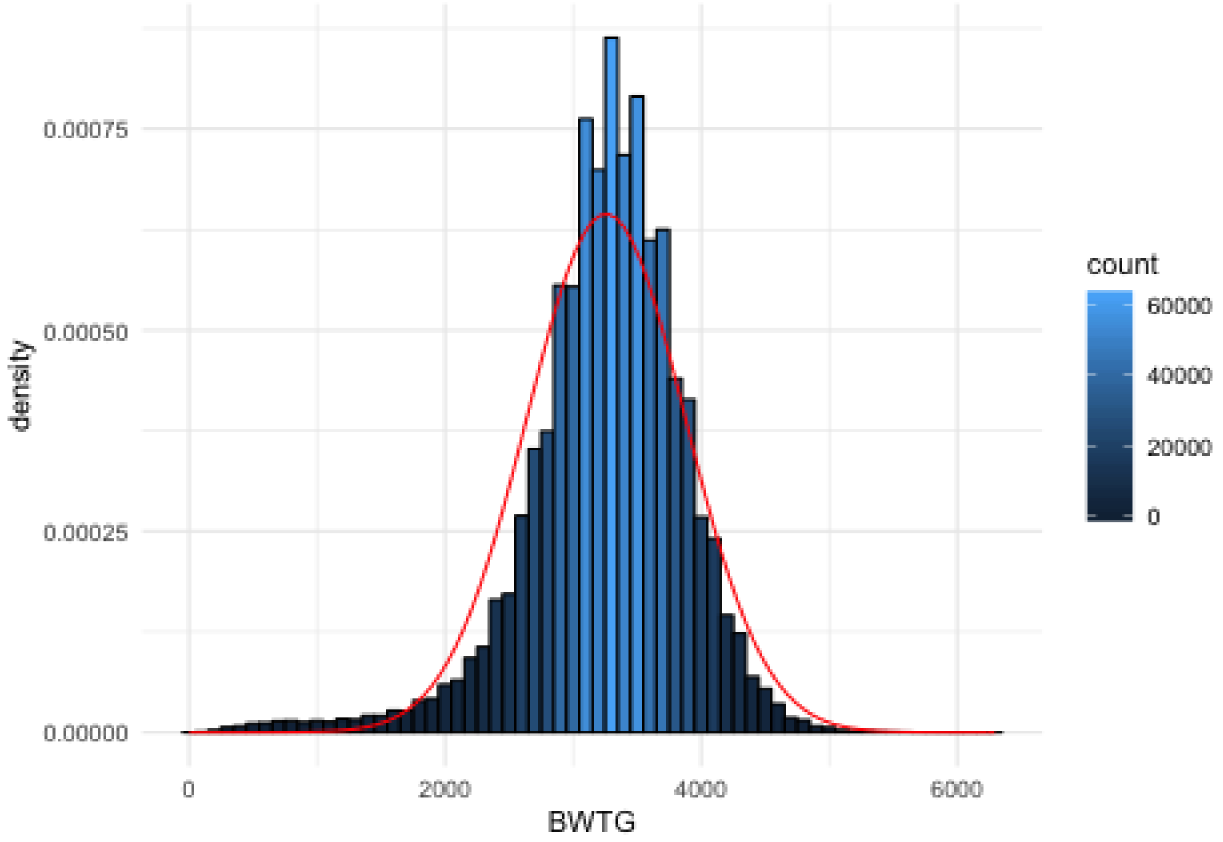
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       4    2948    3310    3258    3640    6294     430
```

```
sd(births$BWTG, na.rm = TRUE)
```

```
## [1] 618.8703
```

```
ggplot(births, aes(x= BWTG))+
  geom_histogram(binwidth=100, colour="black",
                          aes(y=..density.., fill=..count..), position = "stack") +
  stat_function(fun = dnorm, color = "red", args = list(mean = mean(births$BWTG, na.rm = TRUE), sd = sd
  theme_minimal()
```

```
## Warning: Removed 430 rows containing non-finite values (stat_bin).
```



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

```
births = births %>% mutate(
  CIGPN_binary = CIGPN>0,
  CIGFN_binary = CIGFN>0,
  CIGSN_binary = CIGSN>0,
  CIGLN_binary = CIGLN>0
)

births = births %>%
  mutate(total_smoked = CIGFN+CIGSN+CIGLN) %>%
  mutate(smoked_during = total_smoked >0)

mean(births$smoked_during,na.rm=TRUE)
```

```
## [1] 0.100175
```

```
mean(births$CIGPN_binary, na.rm = TRUE)
```

```
## [1] 0.1340957
```

```
mean(births$CIGFN_binar, na.rm = TRUE)
```

```
## [1] 0.09676666
```

```
mean(births$CIGSN_binary,na.rm = TRUE)
```

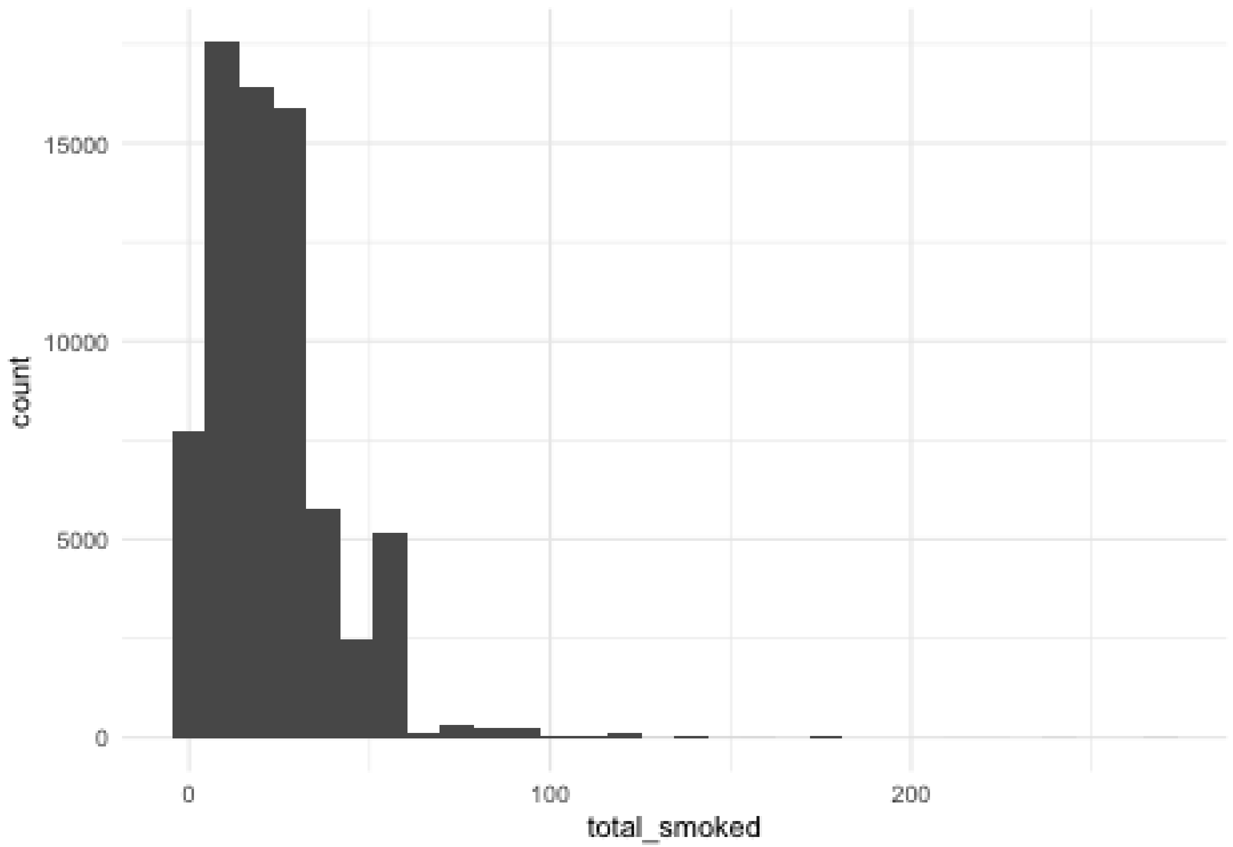
```
## [1] 0.08199303
```

```
mean(births$CIGLN_binary, na.rm = TRUE)
```

```
## [1] 0.07788803
```

```
ggplot(births %>% filter(smoked_during), aes(x = total_smoked))+
  geom_histogram() +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

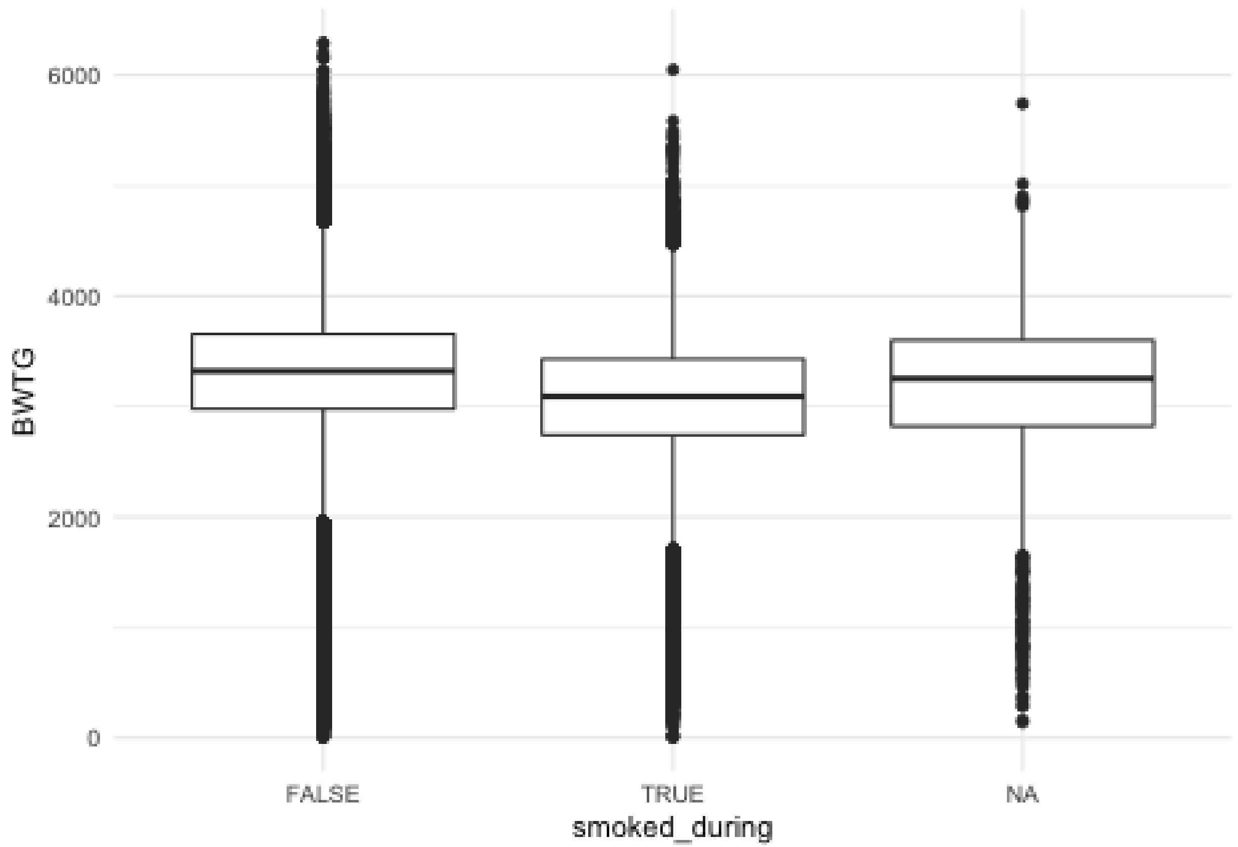


```
mean(births %>% filter(smoked_during) %>% select(total_smoked) %>% unlist())
```

```
## [1] 23.06795
```

```
ggplot(births, aes(x= smoked_during, y = BWTG)) +
  geom_boxplot()+
  theme_minimal()
```

```
## Warning: Removed 430 rows containing non-finite values (stat_boxplot).
```

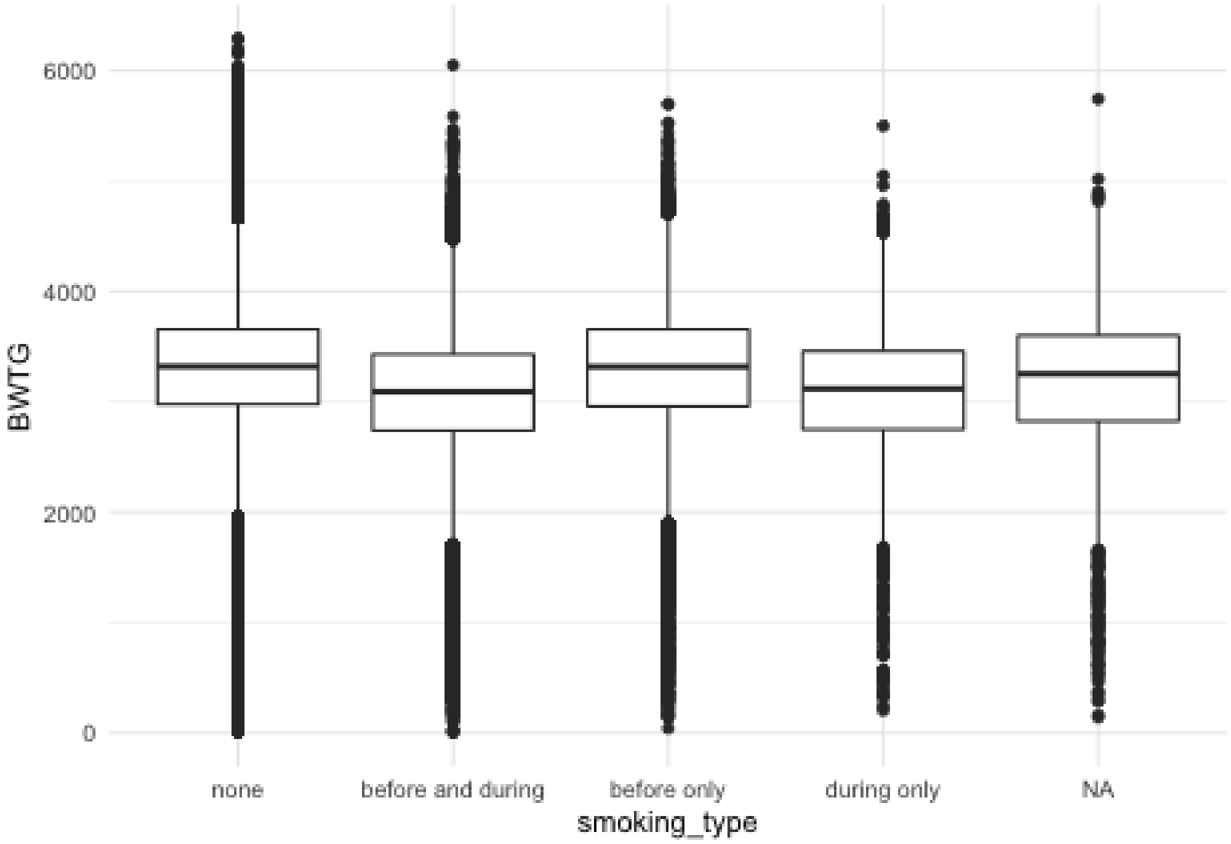


```
lm_smoking_v_non = lm(data=births, BWTG~smoked_during)
summary(lm_smoking_v_non)
```

```
##
## Call:
## lm(formula = BWTG ~ smoked_during, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3276.1  -304.1    40.2   375.9  3012.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3281.064      0.764 4294.52   <2e-16 ***
## smoked_duringTRUE -231.237      2.414  -95.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 614.5 on 718931 degrees of freedom
##   (2759 observations deleted due to missingness)
## Multiple R-squared:  0.0126, Adjusted R-squared:  0.0126
## F-statistic:  9175 on 1 and 718931 DF,  p-value: < 2.2e-16
```

```
births = births %>%
  mutate(smoking_type = ifelse(smoked_during,
      ifelse(CIGPN_binary, "before and during", "during only"),
      ifelse(CIGPN_binary, "before only", "none"))) %>%
  mutate(smoking_type = relevel(as.factor(smoking_type), ref = 4))


ggplot(births, aes(x= smoking_type, y = BWTG)) +
  geom_boxplot()+
  theme_minimal()
```

```
## Warning: Removed 430 rows containing non-finite values (stat_boxplot).
```



```
lm_smoking = lm(data=births, BWTG~smoking_type)
summary(lm_smoking)
```

```
##
## Call:
## lm(formula = BWTG ~ smoking_type, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3276.6  -304.6    40.5   375.4  3012.4
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                   3281.6171     0.7802 4205.990  < 2e-16 ***
## smoking_typebefore and during -232.0954     2.4538  -94.586  < 2e-16 ***
## smoking_typebefore only        -13.4168     3.8483   -3.486  0.00049 ***
## smoking_typeduring only       -223.2977    13.0480  -17.114  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 614.5 on 718904 degrees of freedom
##   (2784 observations deleted due to missingness)
## Multiple R-squared:  0.01262,    Adjusted R-squared:  0.01262
## F-statistic:  3063 on 3 and 718904 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.

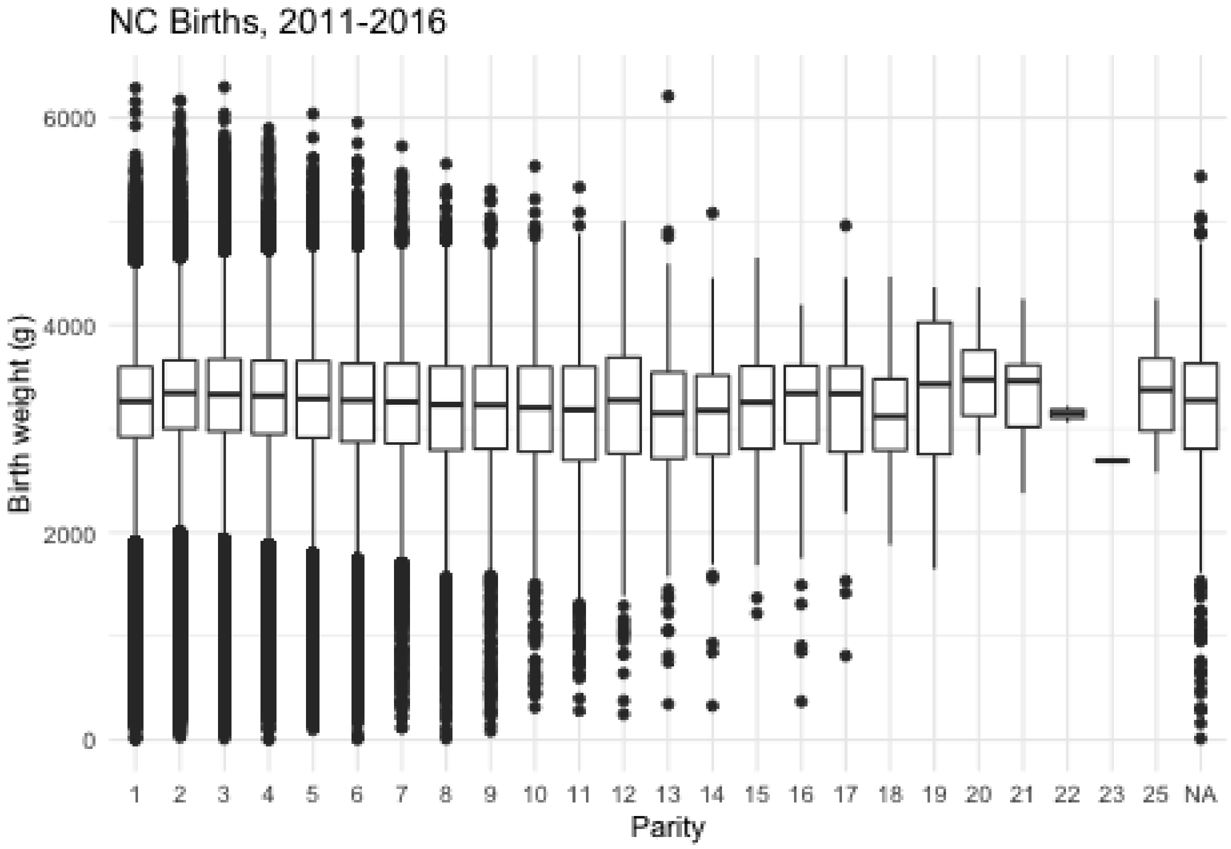
Parity

```r
# Check the parity frequencies
summary(births$PARITY)
```

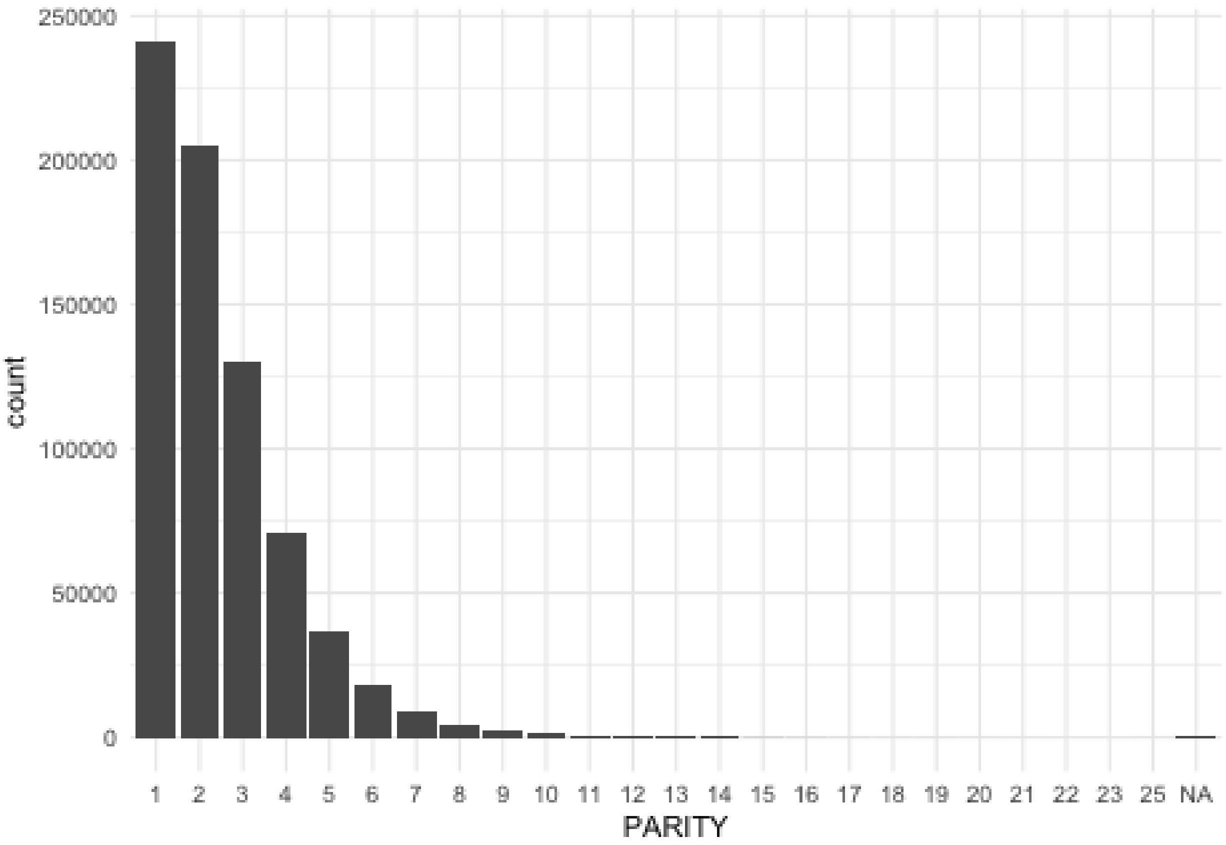
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.000   1.000   2.000   2.465   3.000  25.000     860
```

```r
births$PARITY = as.numeric(births$PARITY)
births = births %>%
  mutate(PARITY_truncated = ifelse(
    PARITY > 4, "5+", PARITY)
  )

births$PARITY = as.factor(births$PARITY)
ggplot(data = births, mapping = aes(x = PARITY, y = BWTG)) +
  geom_boxplot() + xlab("Parity") + ylab("Birth weight (g)") +
  ggtitle("NC Births, 2011-2016") + theme_minimal()
```

```
## Warning: Removed 430 rows containing non-finite values (stat_boxplot).
```



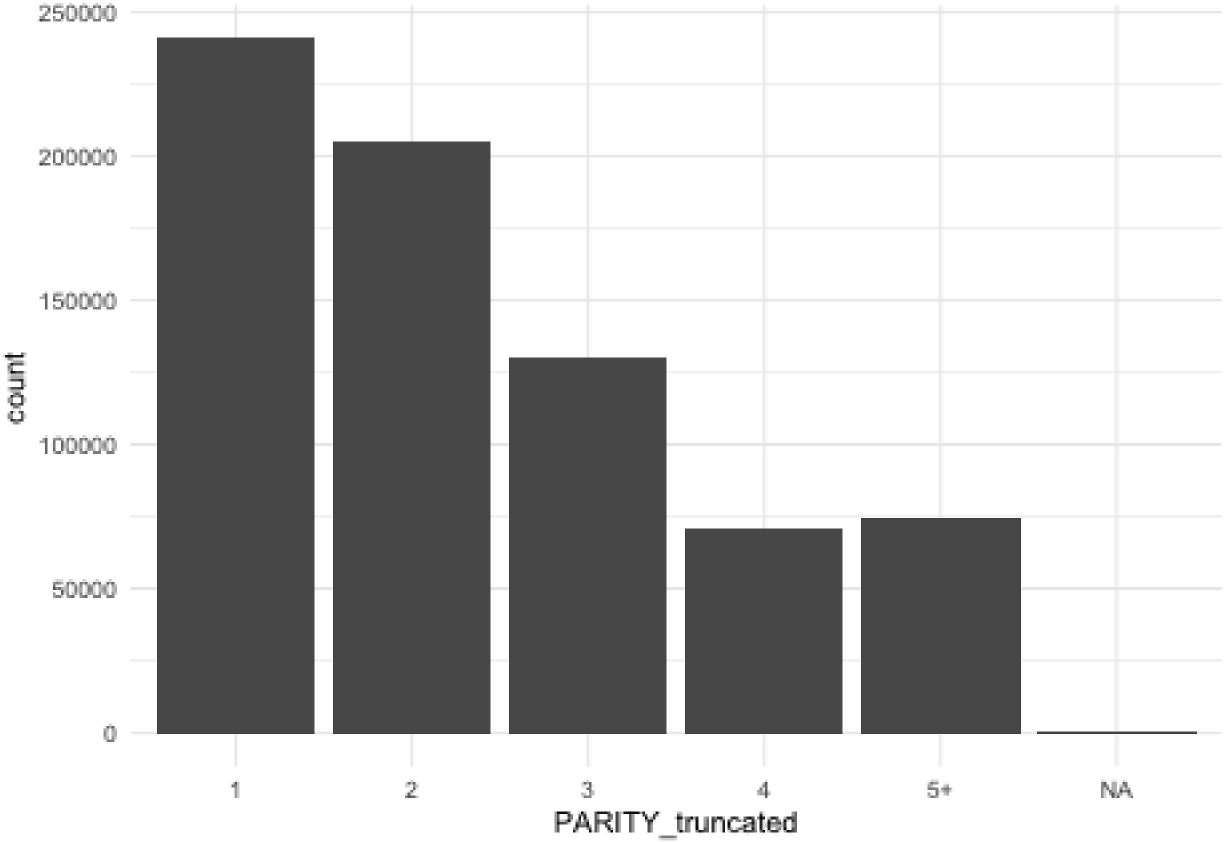
```
ggplot(births, aes(x = PARITY)) +
  stat_count()+
  theme_minimal()
```



```r
# Model without any change, significant up to the low teens
parity_reg = lm(BWTG ~ PARITY, births)
summary(parity_reg)
```

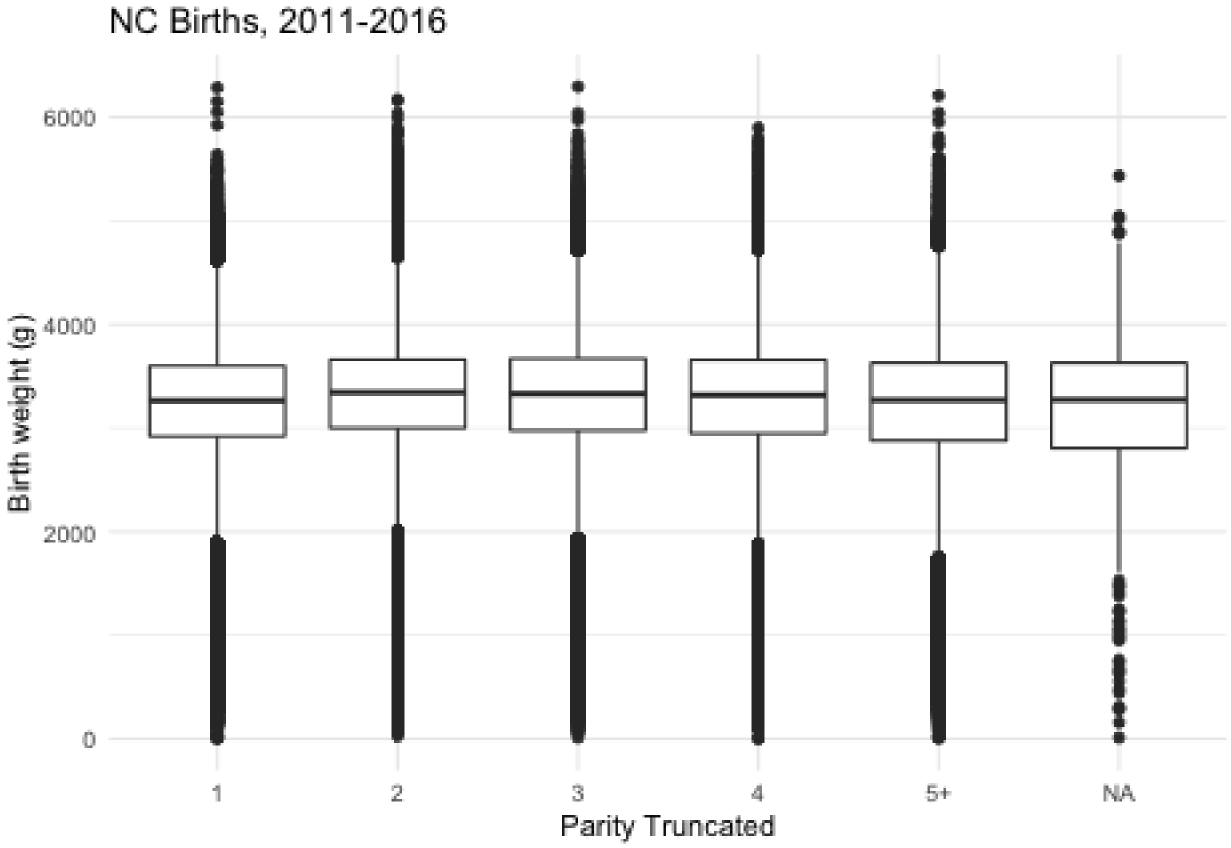
```
##
## Call:
## lm(formula = BWTG ~ PARITY, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3272.9  -303.2    50.9   382.5  3126.8
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 3223.242      1.259 2559.590  < 2e-16 ***
## PARITY2       70.877      1.858   38.155  < 2e-16 ***
## PARITY3       63.680      2.124   29.976  < 2e-16 ***
## PARITY4       40.306      2.645   15.241  < 2e-16 ***
## PARITY5       11.815      3.451    3.424 0.000618 ***
## PARITY6      -11.166      4.751   -2.350 0.018757 *
## PARITY7      -32.906      6.538   -5.033 4.84e-07 ***
## PARITY8      -76.858      9.209   -8.345  < 2e-16 ***
## PARITY9      -70.364     12.447   -5.653 1.58e-08 ***
## PARITY10     -85.904     17.290   -4.968 6.75e-07 ***
## PARITY11    -126.567     22.324   -5.670 1.43e-08 ***
## PARITY12     -41.074     30.094   -1.365 0.172306
## PARITY13    -140.997     37.611   -3.749 0.000178 ***
## PARITY14    -157.146     55.261   -2.844 0.004459 **
## PARITY15     -20.675     62.728   -0.330 0.741706
## PARITY16    -158.545     82.550   -1.921 0.054782 .
## PARITY17     -64.515    107.530   -0.600 0.548530
## PARITY18     -61.242    145.592   -0.421 0.674019
## PARITY19      62.758    165.085    0.380 0.703829
## PARITY20     250.258    195.329    1.281 0.200120
## PARITY21     112.901    233.462    0.484 0.628673
## PARITY22     -80.242    356.616   -0.225 0.821972
## PARITY23    -530.242    617.674   -0.858 0.390645
## PARITY25     162.758    218.384    0.745 0.456100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 617.7 on 720622 degrees of freedom
##   (1046 observations deleted due to missingness)
## Multiple R-squared:  0.003304,	Adjusted R-squared:  0.003272 
## F-statistic: 103.9 on 23 and 720622 DF,  p-value: < 2.2e-16
```

```
# Rerun EDA with truncated data
births$PARITY_truncated = as.factor(births$PARITY_truncated)
ggplot(births, aes(x = PARITY_truncated)) +
  stat_count() +
  theme_minimal()
```



```
ggplot(data = births, mapping = aes(x = PARITY_truncated, y = BWTG)) +
  geom_boxplot() + xlab("Parity Truncated") + ylab("Birth weight (g)") +
  ggtitle("NC Births, 2011-2016") +
  theme_minimal()
```

```
## Warning: Removed 430 rows containing non-finite values (stat_boxplot).
```



```
parity_truncated_reg = lm(BWTG ~ PARITY_truncated, births)
summary(parity_truncated_reg)
```

```
##
## Call:
## lm(formula = BWTG ~ PARITY_truncated, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3272.9  -303.2    48.7   383.1  3061.8
##
## Coefficients:
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         3223.242      1.259 2559.251  < 2e-16 ***
## PARITY_truncated2     70.877      1.858   38.150  < 2e-16 ***
## PARITY_truncated3     63.680      2.125   29.972  < 2e-16 ***
## PARITY_truncated4     40.306      2.645   15.239  < 2e-16 ***
## PARITY_truncated5+   -11.939      2.589   -4.612 3.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

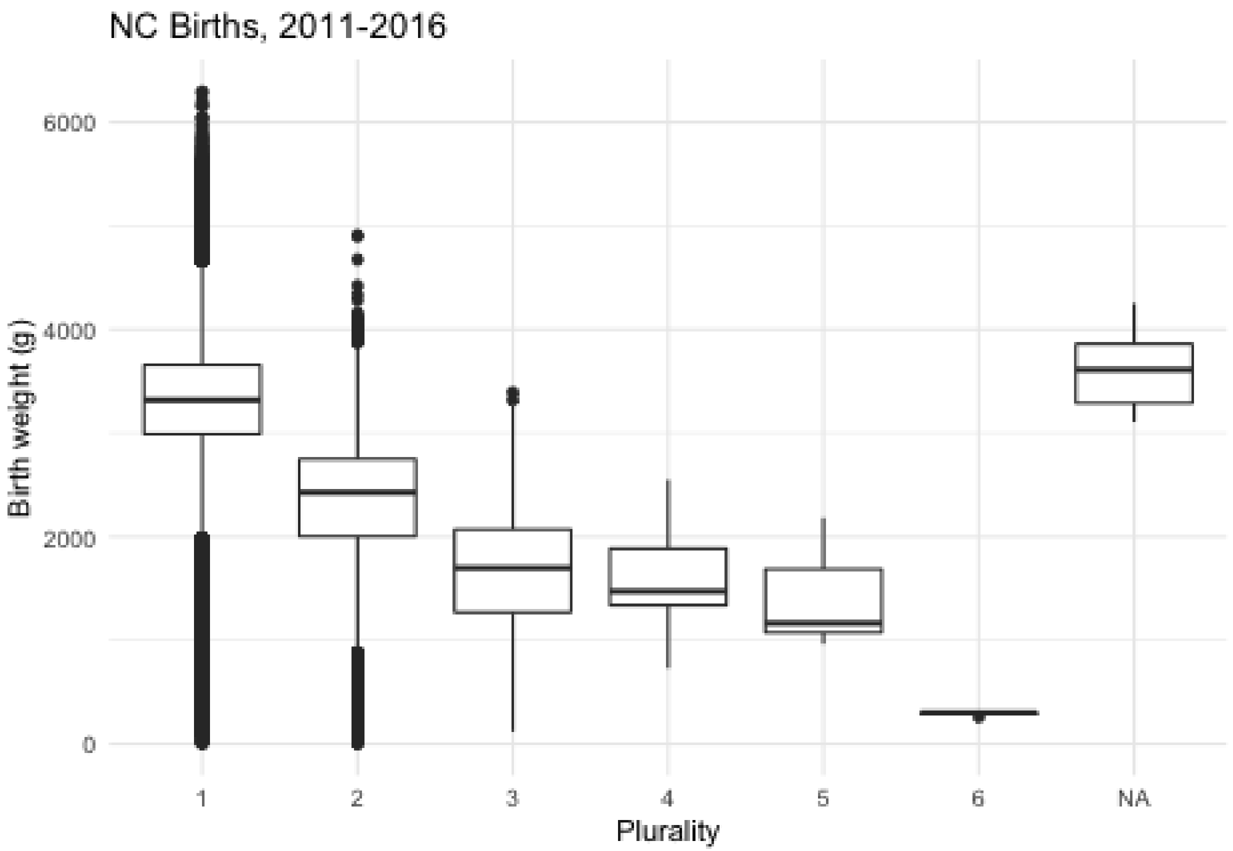
```
##
## Residual standard error: 617.8 on 720641 degrees of freedom
##   (1046 observations deleted due to missingness)
## Multiple R-squared:  0.003014,   Adjusted R-squared:  0.003008
## F-statistic: 544.6 on 4 and 720641 DF,  p-value: < 2.2e-16
```

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.

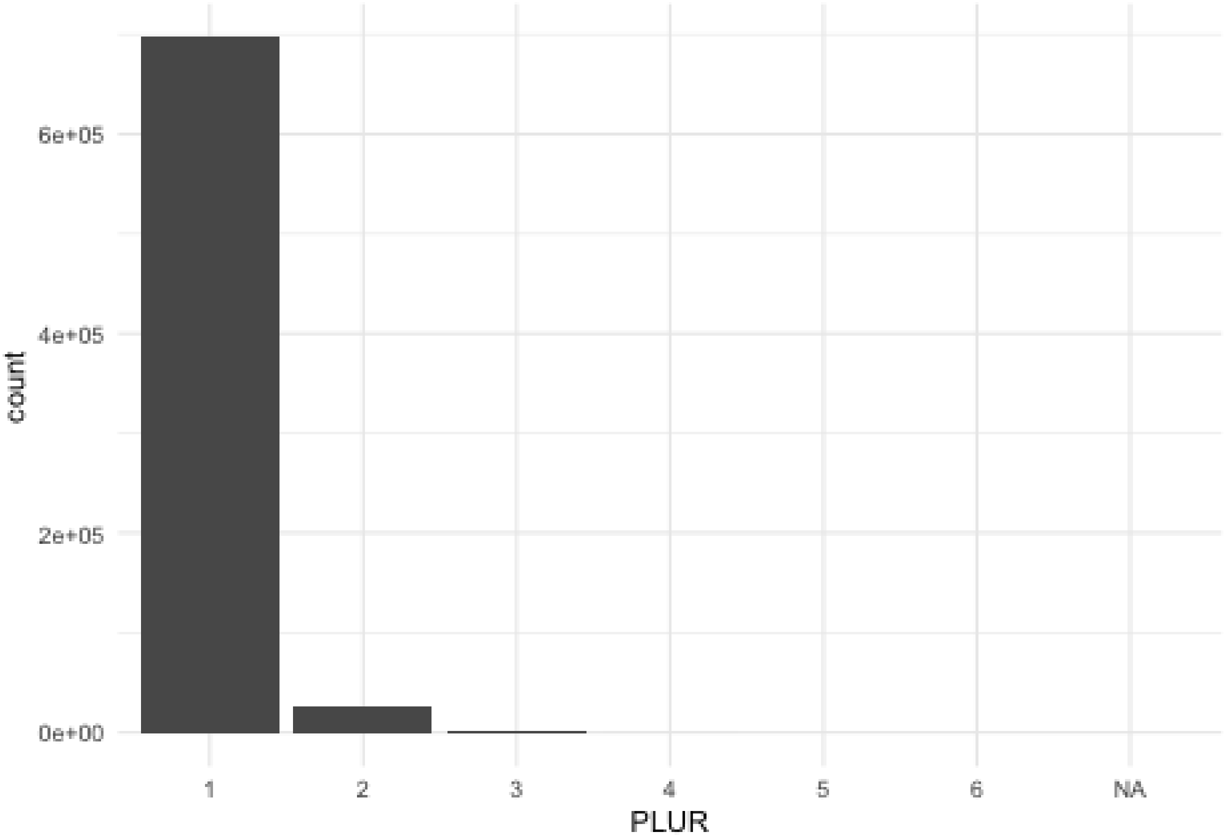
Plurality

```
# Check the plurality frequencies
births$PLUR = as.numeric(births$PLUR)

births = births %>%
  mutate(PLUR_truncated = ifelse(PLUR > 2, "3+", PLUR))

births$PLUR = as.factor(births$PLUR)
ggplot(data = births, mapping = aes(x = PLUR, y = BWTG)) +
  geom_boxplot() +
  xlab("Plurality") +
  ylab("Birth weight (g)") +
  ggtitle("NC Births, 2011-2016") +
  theme_minimal()
```

```
## Warning: Removed 430 rows containing non-finite values (stat_boxplot).
```



```r
ggplot(births, aes(x = PLUR)) +
  stat_count() +
  theme_minimal()
```

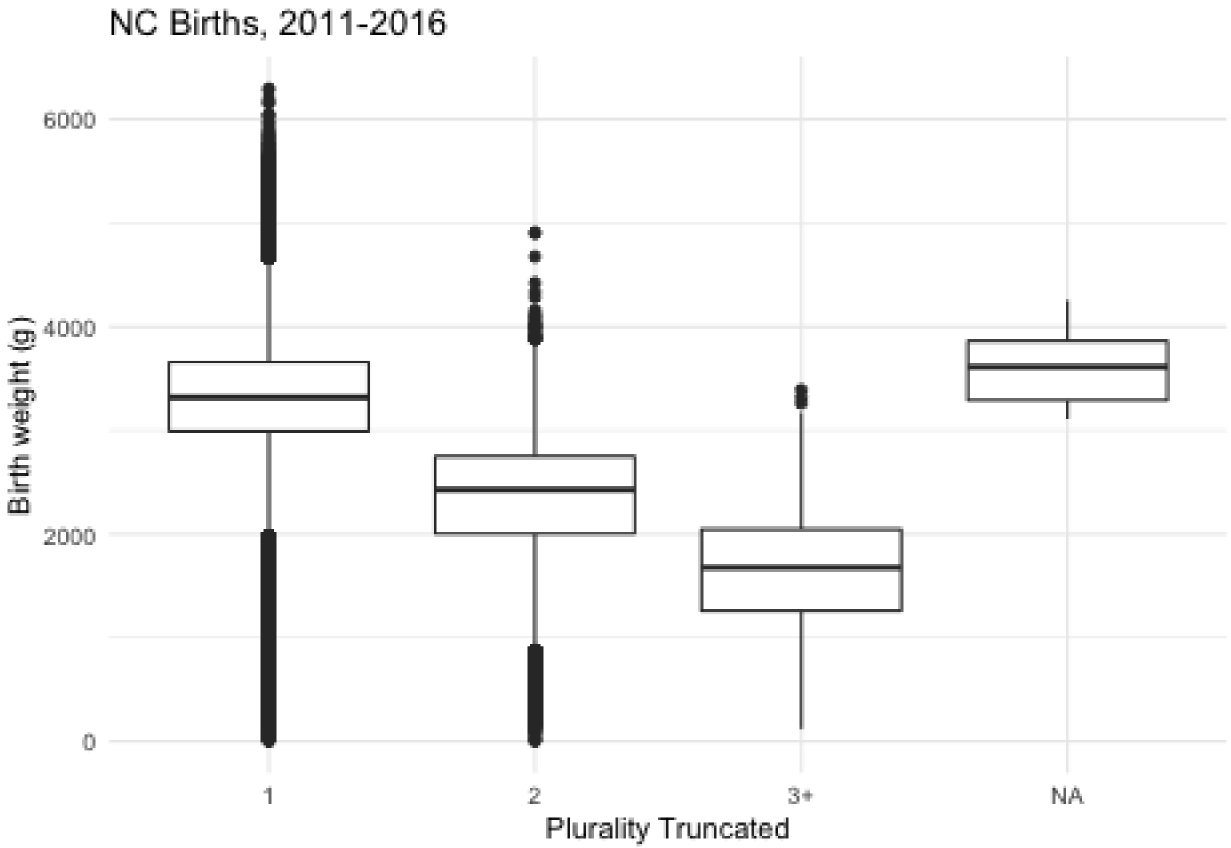


```
plurality_reg = lm(BWTG ~ PLUR, births)
summary(plurality_reg)
```

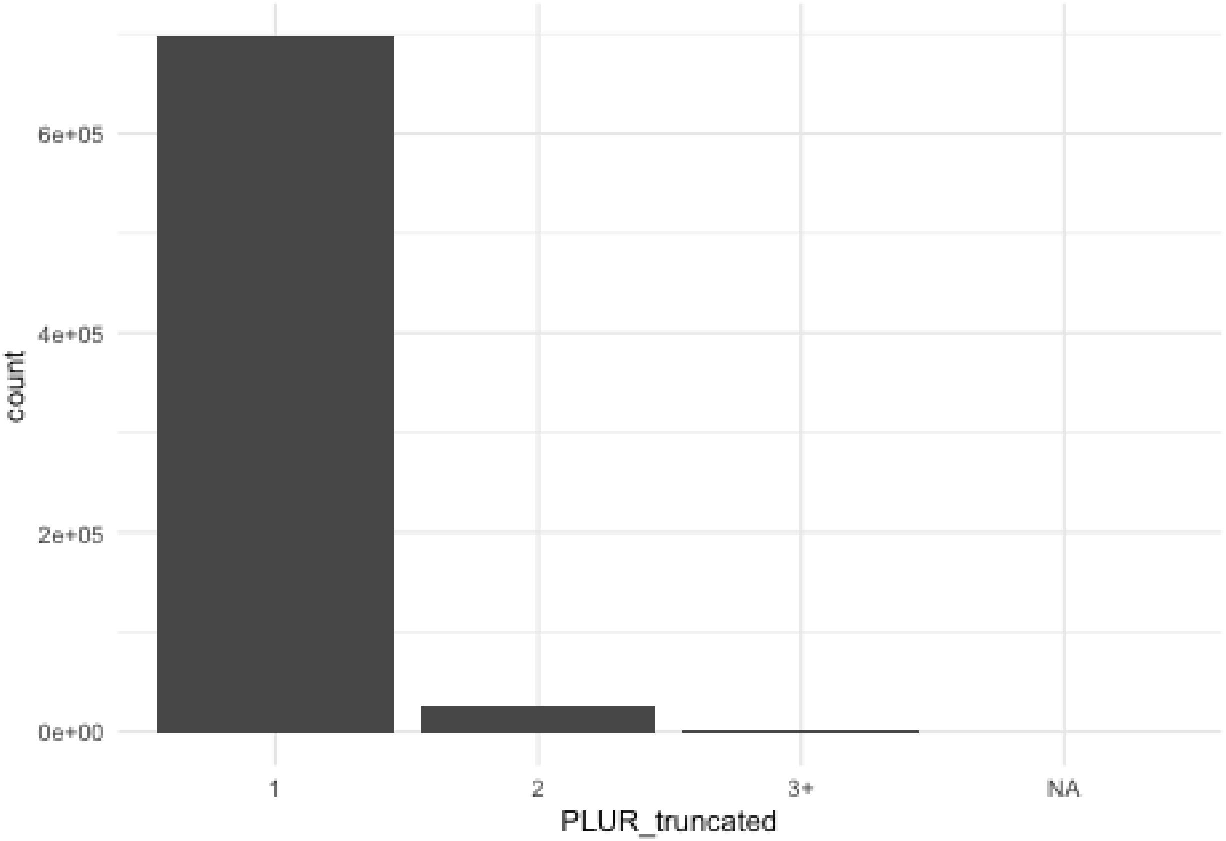
```
##
## Call:
## lm(formula = BWTG ~ PLUR, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3287.1  -302.1    28.9   364.9  3001.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3292.1374     0.7085 4646.93   <2e-16 ***
## PLUR2       -969.8739     3.8495 -251.95   <2e-16 ***
## PLUR3      -1632.6058    21.4519  -76.11   <2e-16 ***
## PLUR4      -1767.3874   132.1681  -13.37   <2e-16 ***
## PLUR5      -1899.7374   186.9125  -10.16   <2e-16 ***
## PLUR6      -2999.9707   241.3023  -12.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 591.1 on 721251 degrees of freedom
##   (435 observations deleted due to missingness)
## Multiple R-squared:  0.08785,    Adjusted R-squared:  0.08784
## F-statistic: 1.389e+04 on 5 and 721251 DF,  p-value: < 2.2e-16
```

```
births$PLUR_truncated = as.factor(births$PLUR_truncated)

ggplot(data = births, mapping = aes(x = PLUR_truncated, y = BWTG)) +
  geom_boxplot() +
  xlab("Plurality Truncated") +
  ylab("Birth weight (g)") +
  ggtitle("NC Births, 2011-2016") +
  theme_minimal()
```

```
## Warning: Removed 430 rows containing non-finite values (stat_boxplot).
```



```
ggplot(births, aes(x = PLUR_truncated)) +
  stat_count() +
  theme_minimal()
```



```
plurality_reg_truncated = lm(BWTG ~ PLUR_truncated, births)
summary(plurality_reg_truncated)
```

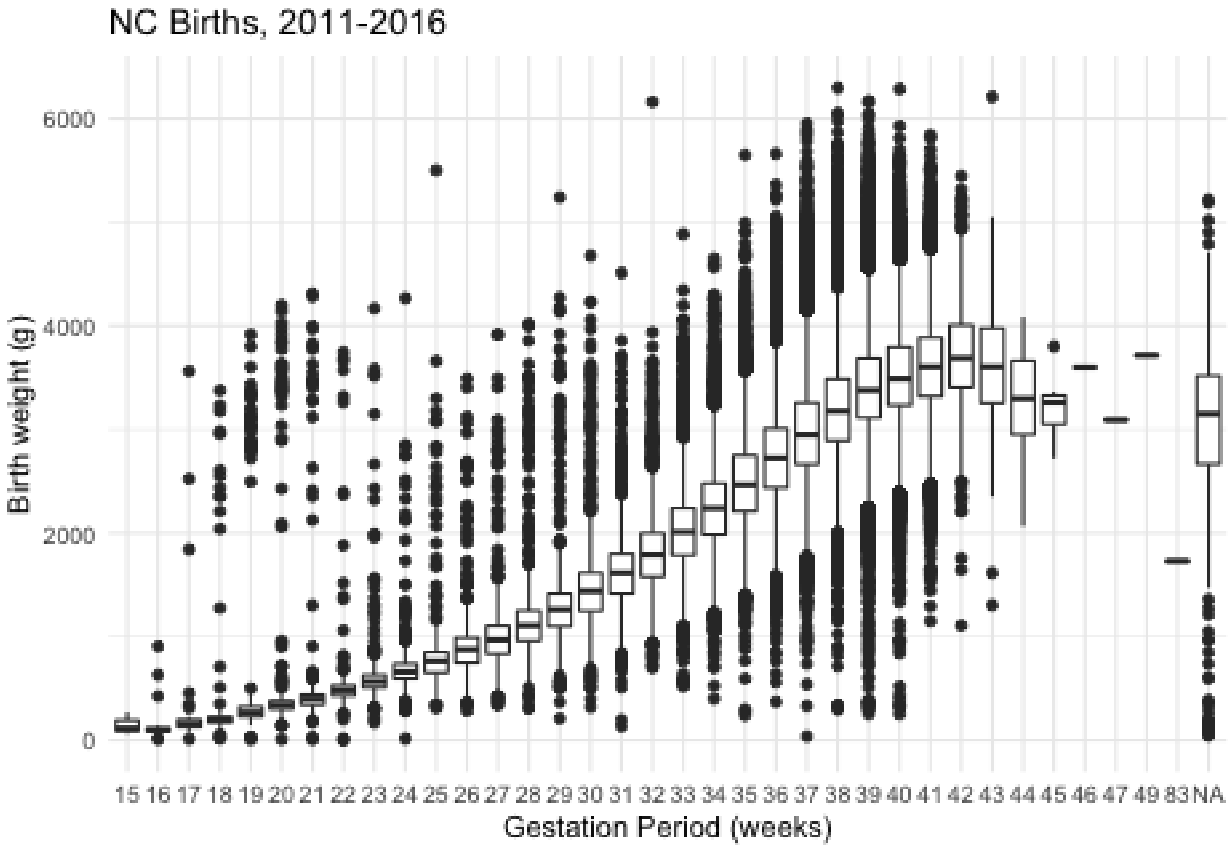
```
##
## Call:
## lm(formula = BWTG ~ PLUR_truncated, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3287.1  -302.1    28.9   364.9  3001.9
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3292.1374     0.7085  4646.8   <2e-16 ***
## PLUR_truncated2   -969.8739     3.8496  -251.9   <2e-16 ***
## PLUR_truncated3+ -1649.6550    20.9622   -78.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 591.1 on 721254 degrees of freedom
##   (435 observations deleted due to missingness)
## Multiple R-squared:  0.0878, Adjusted R-squared:  0.0878
## F-statistic: 3.471e+04 on 2 and 721254 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

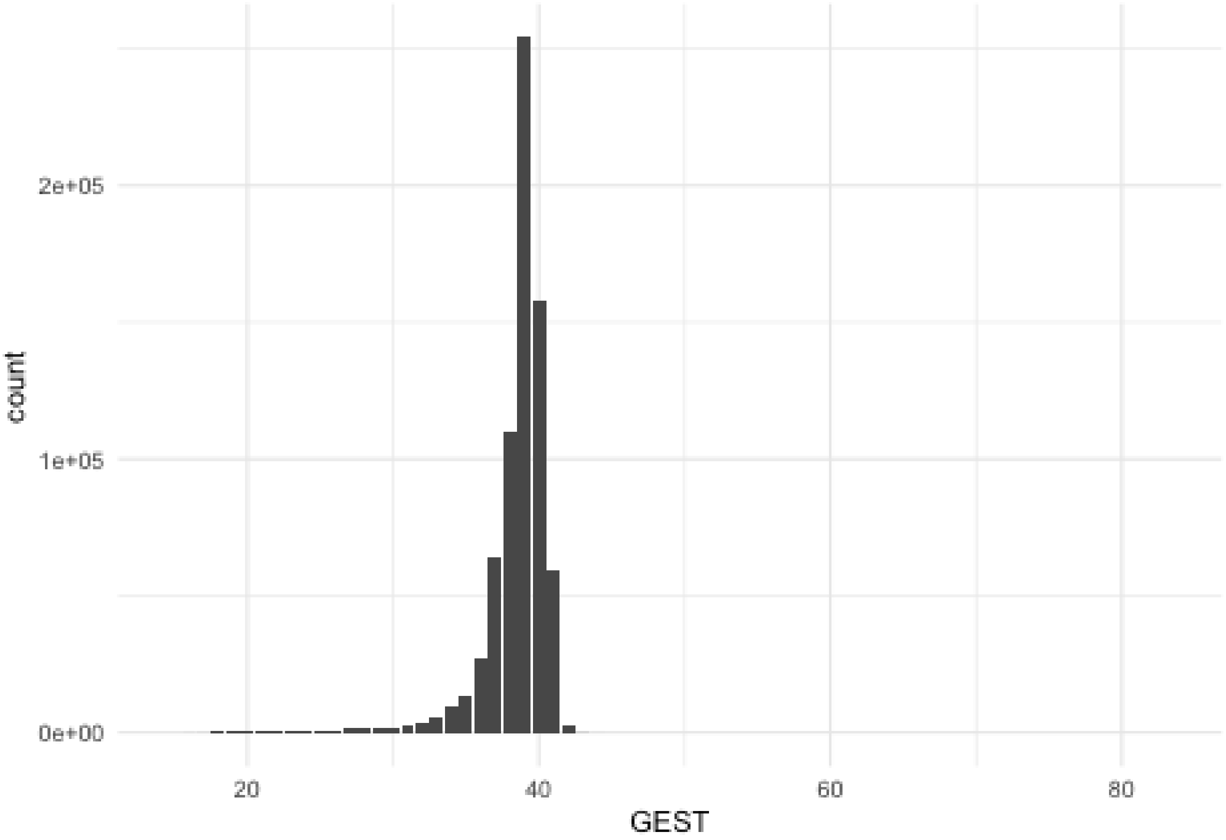
```
ggplot(data = births, mapping = aes(x = as.factor(GEST), y = BWTG)) +
  xlab("Gestation Period (weeks)") +
  ylab("Birth weight (g)") +
  geom_boxplot() +
  ggtitle("NC Births, 2011-2016") +
  theme_minimal()
```

```
## Warning: Removed 430 rows containing non-finite values (stat_boxplot).
```



```
ggplot(births, aes(x = GEST)) +
  stat_count() +
  theme_minimal()
```

```
## Warning: Removed 539 rows containing non-finite values (stat_count).
```



```
gest_reg = lm(BWTG ~ GEST, births)
summary(gest_reg)
```

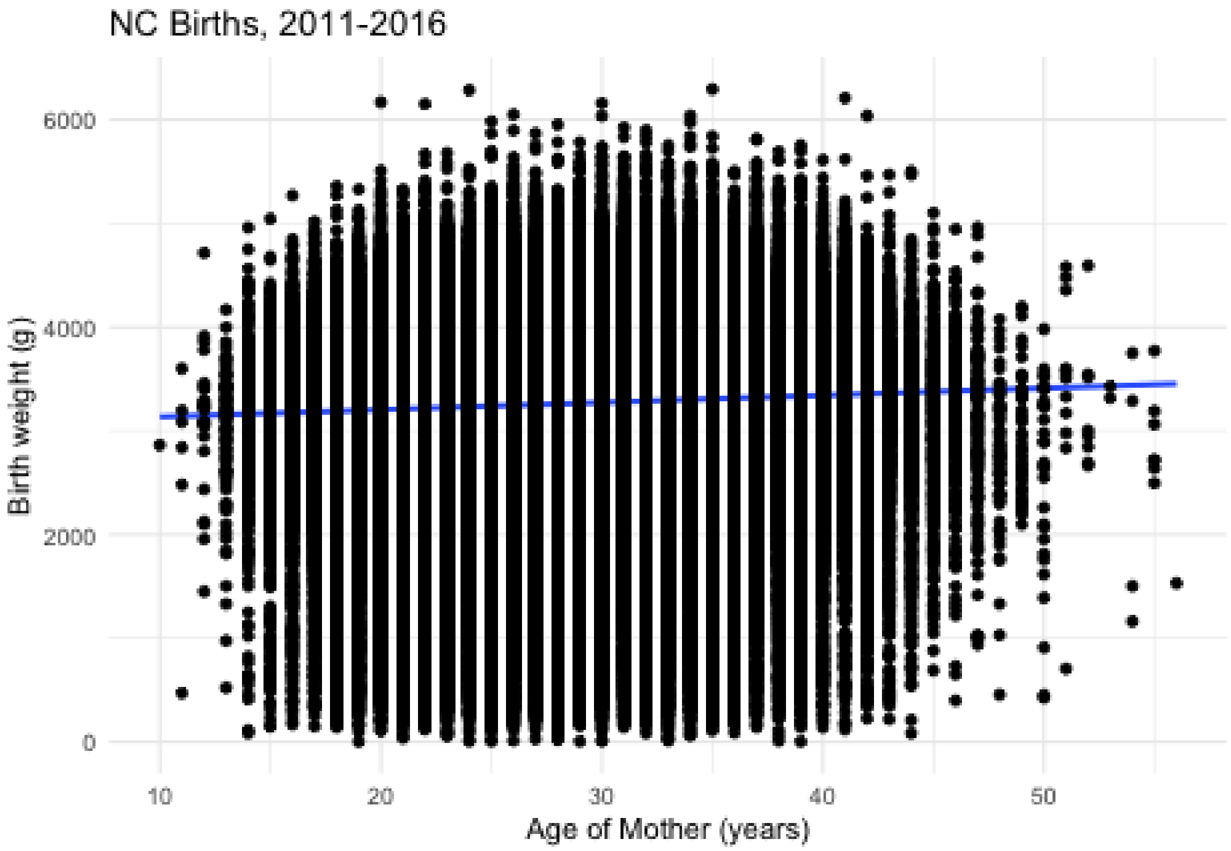
```
##
## Call:
## lm(formula = BWTG ~ GEST, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9797.7  -299.6   -20.5   276.2  4752.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3898.2669     8.8696  -439.5   <2e-16 ***
## GEST          185.8433     0.2299   808.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 448.1 on 720801 degrees of freedom
##   (889 observations deleted due to missingness)
## Multiple R-squared:  0.4754, Adjusted R-squared:  0.4754
## F-statistic: 6.532e+05 on 1 and 720801 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

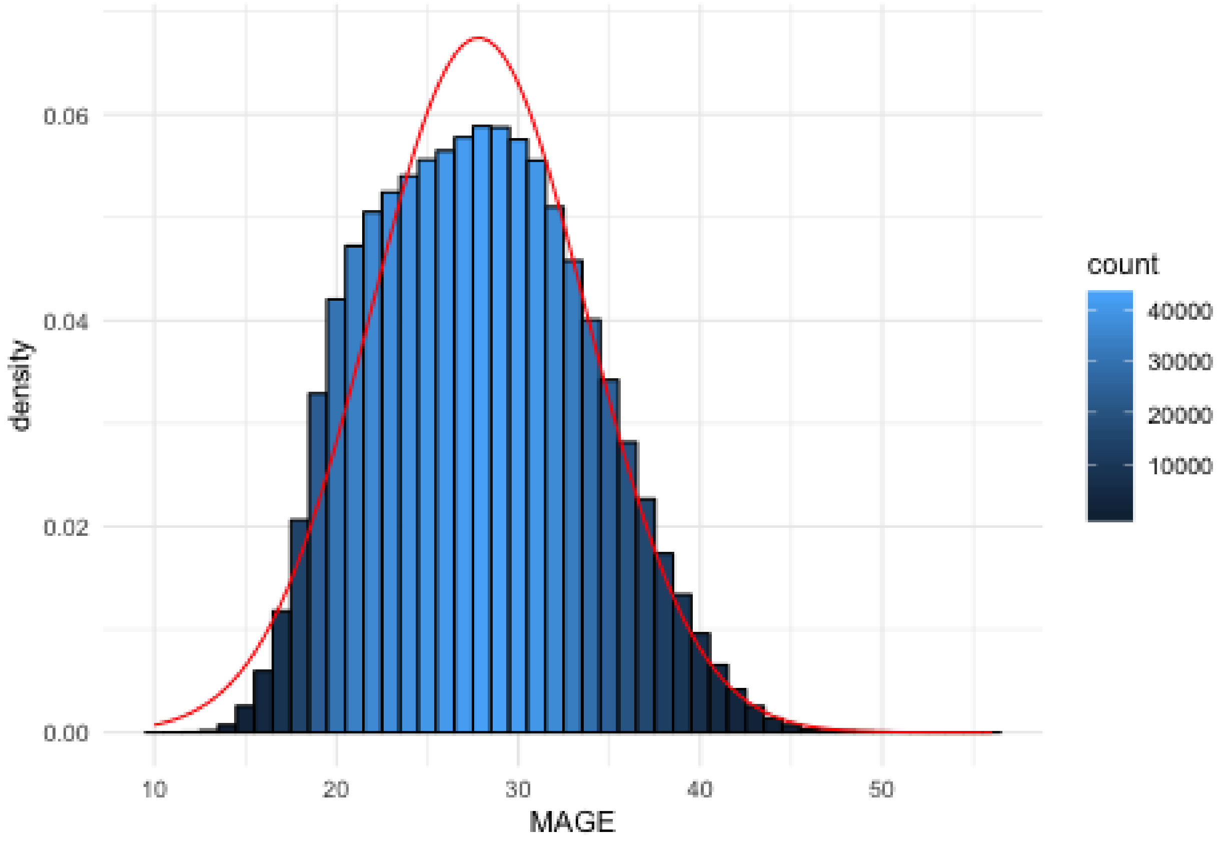
```
ggplot(data = births, mapping = aes(x = MAGE, y = BWTG)) +
  xlab("Age of Mother (years)") +
  ylab("Birth weight (g)") +
  geom_smooth(method='lm', na.rm = TRUE) +
  geom_point() +
  ggtitle("NC Births, 2011-2016") +
  theme_minimal()
```

```
## Warning: Removed 457 rows containing missing values (geom_point).
```



```
ggplot(births, aes(x = MAGE)) +
  geom_histogram(binwidth=1, colour="black",
                        aes(y=..density.., fill=..count..), position = "stack") +
  stat_function(fun = dnorm, color = "red", args = list(mean = mean(births$MAGE, na.rm = TRUE), sd = sd
  theme_minimal()
```

```
## Warning: Removed 27 rows containing non-finite values (stat_bin).
```



```
mage_reg = lm(BWTG ~ MAGE, births)
summary(mage_reg)
```

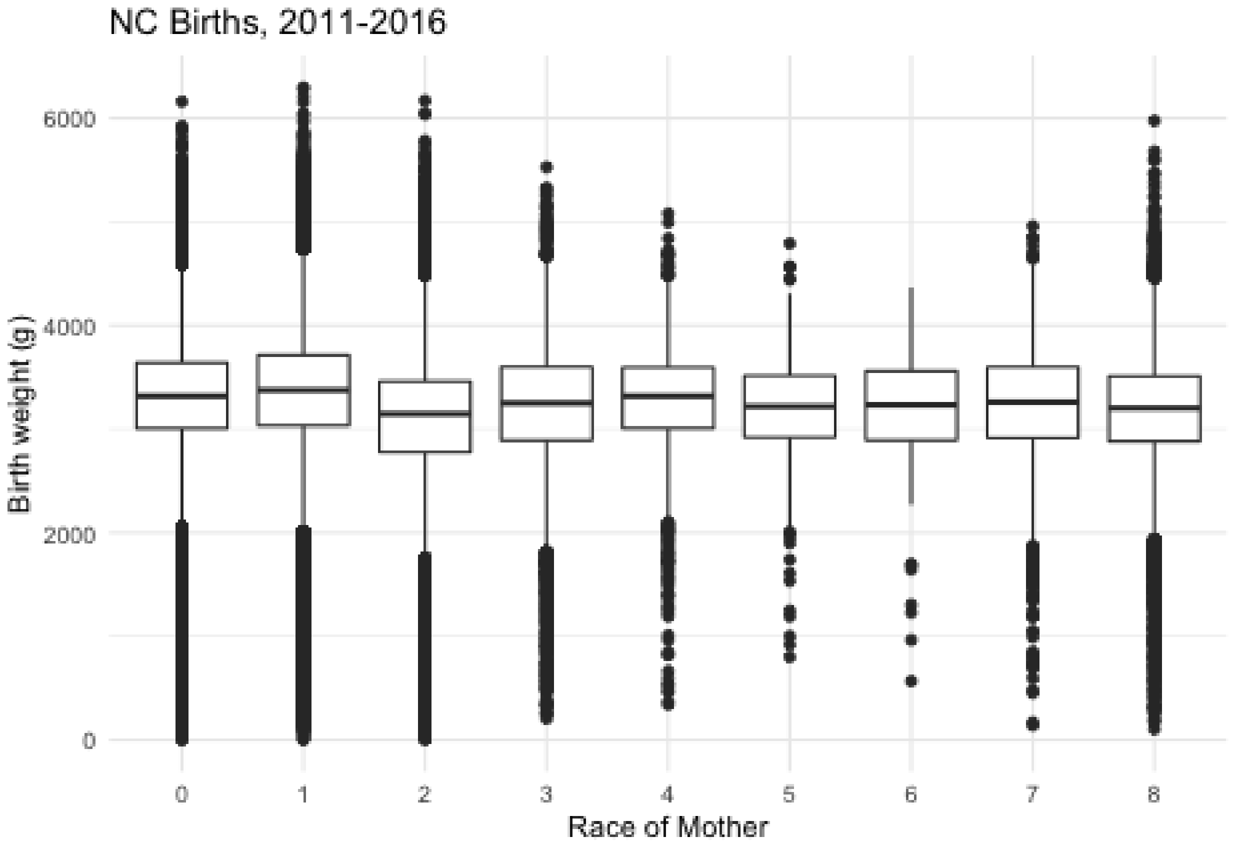
```
##
## Call:
## lm(formula = BWTG ~ MAGE, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3330.1  -302.3    49.7   383.5  3053.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3064.715      3.496  876.58   <2e-16 ***
## MAGE           6.932      0.123   56.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 617.5 on 721233 degrees of freedom
##   (457 observations deleted due to missingness)
## Multiple R-squared:  0.004388,	Adjusted R-squared:  0.004386
## F-statistic:  3178 on 1 and 721233 DF,  p-value: < 2.2e-16
```

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

```
births$MRACER = as.factor(births$MRACER)
ggplot(data = births, mapping = aes(x = MRACER, y = BWTG)) +
  xlab("Race of Mother") +
  ylab("Birth weight (g)") +
  geom_boxplot() +
  ggtitle("NC Births, 2011-2016") +
  theme_minimal()
```

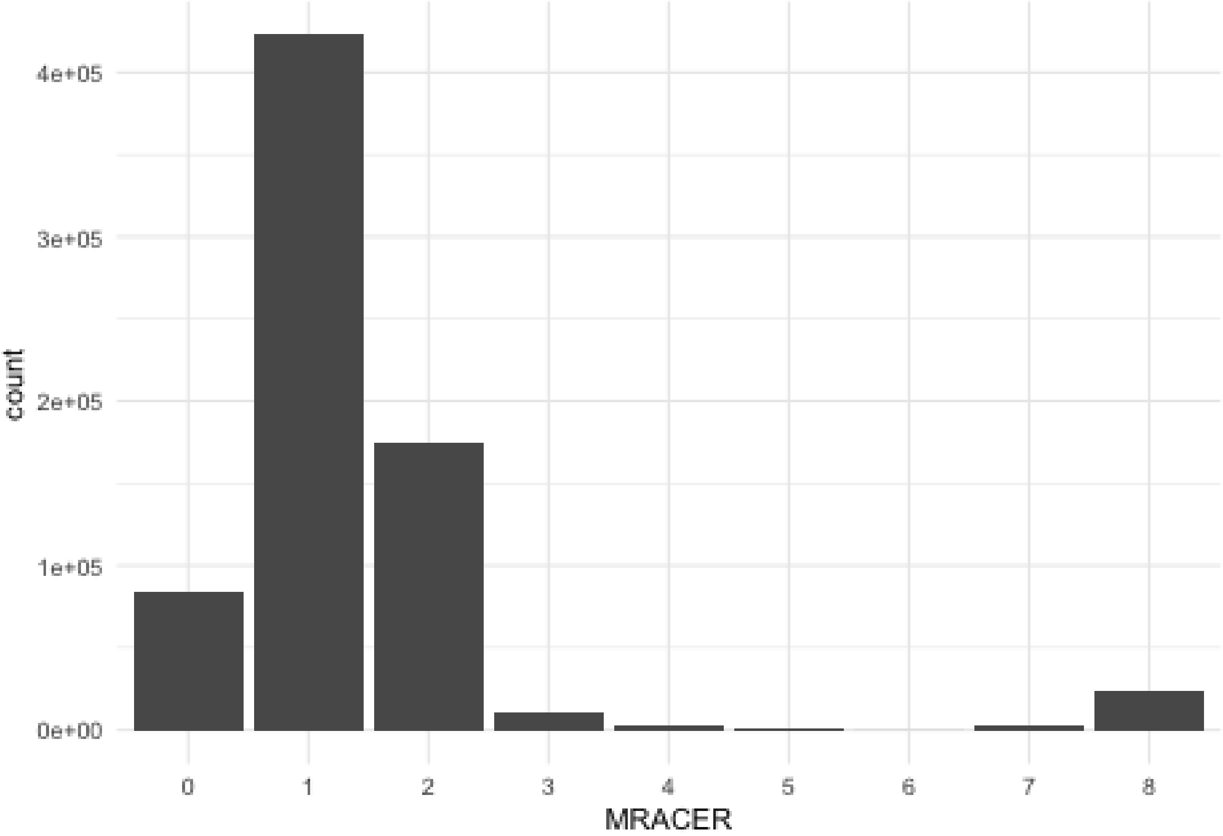
```
## Warning: Removed 430 rows containing non-finite values (stat_boxplot).
```



```
bw_race_avgs = births %>%
  group_by(MRACER) %>%
  summarize(mweight = mean(BWTG, na.rm = TRUE), freq_percent = n()/nrow(births))
bw_race_avgs
```

```
## # A tibble: 9 x 3
##   MRACER mweight freq_percent
##   <fct>    <dbl>        <dbl>
## 1 0        3292.     0.117
## 2 1        3335.     0.586
## 3 2        3069.     0.243
## 4 3        3196.     0.0141
## 5 4        3285.     0.00463
## 6 5        3193.     0.000806
## 7 6        3148      0.000132
## 8 7        3218.     0.00294
## 9 8        3169.     0.0322
```

```
ggplot(births, aes(x = MRACER)) +
  stat_count() +
  theme_minimal()
```



```
mage_reg = lm(BWTG ~ MRACER, births)
summary(mage_reg)
```

```
##
## Call:
## lm(formula = BWTG ~ MRACER, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3331.1  -301.3    48.7   378.9  3100.7
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 3291.649      2.095 1571.411  < 2e-16 ***
## MRACER1       43.495      2.295   18.956  < 2e-16 ***
## MRACER2     -222.359      2.550  -87.191  < 2e-16 ***
## MRACER3      -95.772      6.394  -14.978  < 2e-16 ***
## MRACER4       -7.146     10.741   -0.665 0.505820
## MRACER5      -98.199     25.315   -3.879 0.000105 ***
## MRACER6     -143.649     62.479   -2.299 0.021496 *
## MRACER7      -73.792     13.377   -5.516 3.46e-08 ***
## MRACER8     -122.502      4.511  -27.159  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 608.6 on 721253 degrees of freedom
##   (430 observations deleted due to missingness)
## Multiple R-squared:  0.03285,    Adjusted R-squared:  0.03284
## F-statistic:  3062 on 8 and 721253 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.

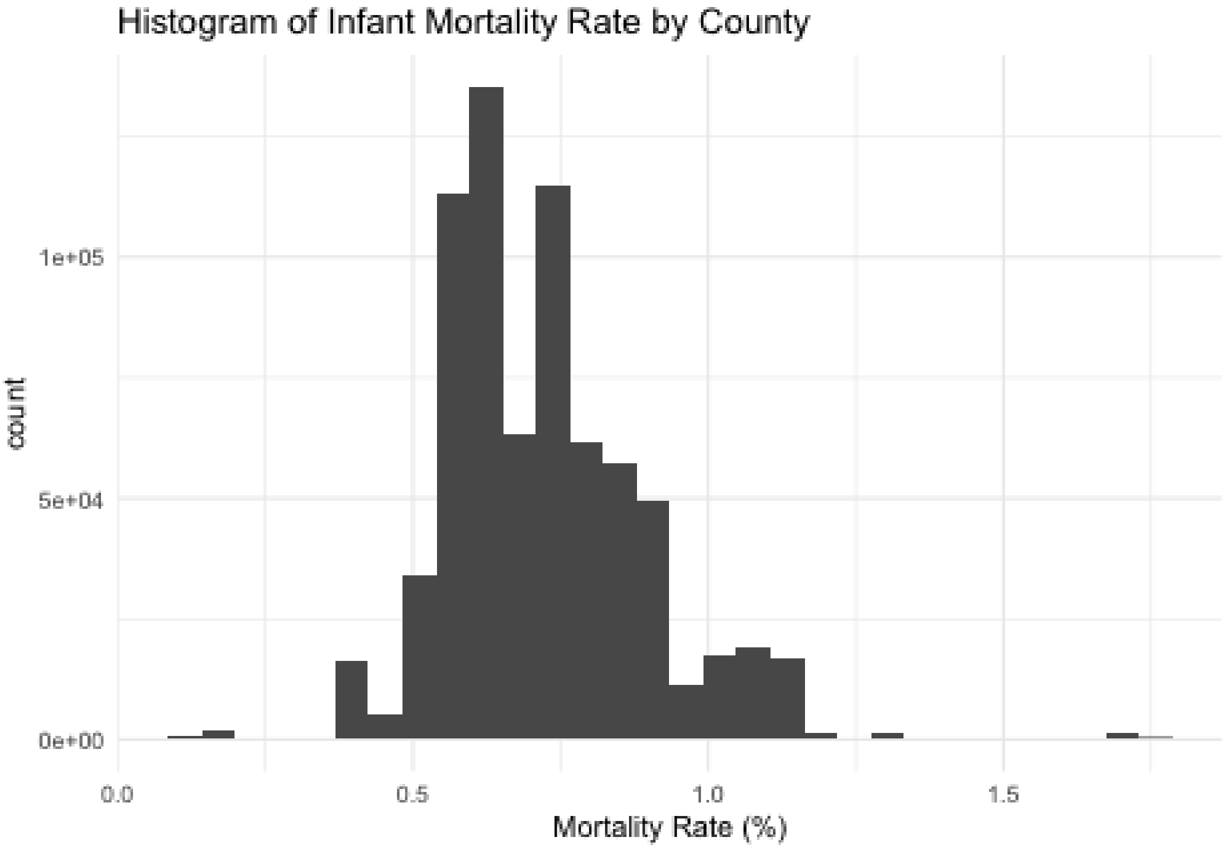
County / Socioeconomic Status

```
#calculate infant mortality by county, to use as a proxy for socioeconomic status
deaths_by_county = deaths %>%
  group_by(cores) %>%
  summarize(n_deaths = n()) %>%
  rename(CORES=cores)

births_by_county = births %>%
  group_by(CORES) %>%
  summarize(n_births = n())

infant_mortality = merge(deaths_by_county, births_by_county, by = "CORES") %>%
  mutate(mortality = n_deaths/n_births*100) %>% #note: mortality rate is as percentage
  select(-n_births, -n_deaths)

births = merge(births, infant_mortality, by = "CORES")

summary(births$mortality)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1166  0.6149  0.7009  0.7203  0.7958  1.7606
```

```
ggplot(births, aes(x= mortality))+
  geom_histogram() +
  xlab("Mortality Rate (%)")+
  theme_minimal() +
  ggtitle("Histogram of Infant Mortality Rate by County")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

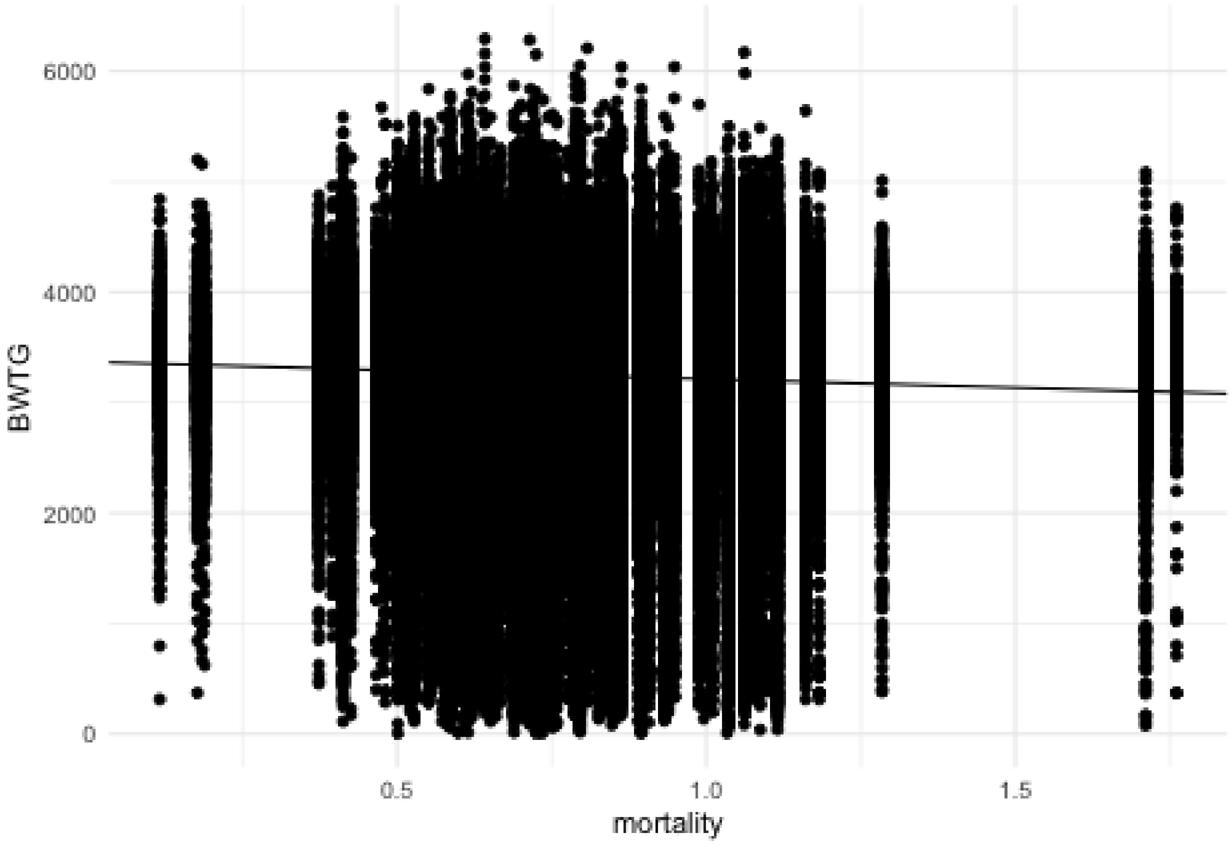


```
bwtg_vs_mortality = lm(data = births, BWTG~mortality)
summary(bwtg_vs_mortality)
```

```
##
## Call:
## lm(formula = BWTG ~ mortality, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3286.9  -303.6    45.3   382.9  3026.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3370.843      3.187 1057.66   <2e-16 ***
## mortality   -157.360      4.308  -36.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 618.3 on 721260 degrees of freedom
##   (430 observations deleted due to missingness)
## Multiple R-squared:  0.001846,   Adjusted R-squared:  0.001845
## F-statistic:  1334 on 1 and 721260 DF,  p-value: < 2.2e-16
```

```
ggplot(births, aes(x = mortality, y = BWTG))+
  geom_point() +
  geom_abline(slope = bwtg_vs_mortality$coefficients[[2]], intercept = bwtg_vs_mortality$coefficients[[
  theme_minimal()
```

```
## Warning: Removed 430 rows containing missing values (geom_point).
```



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.

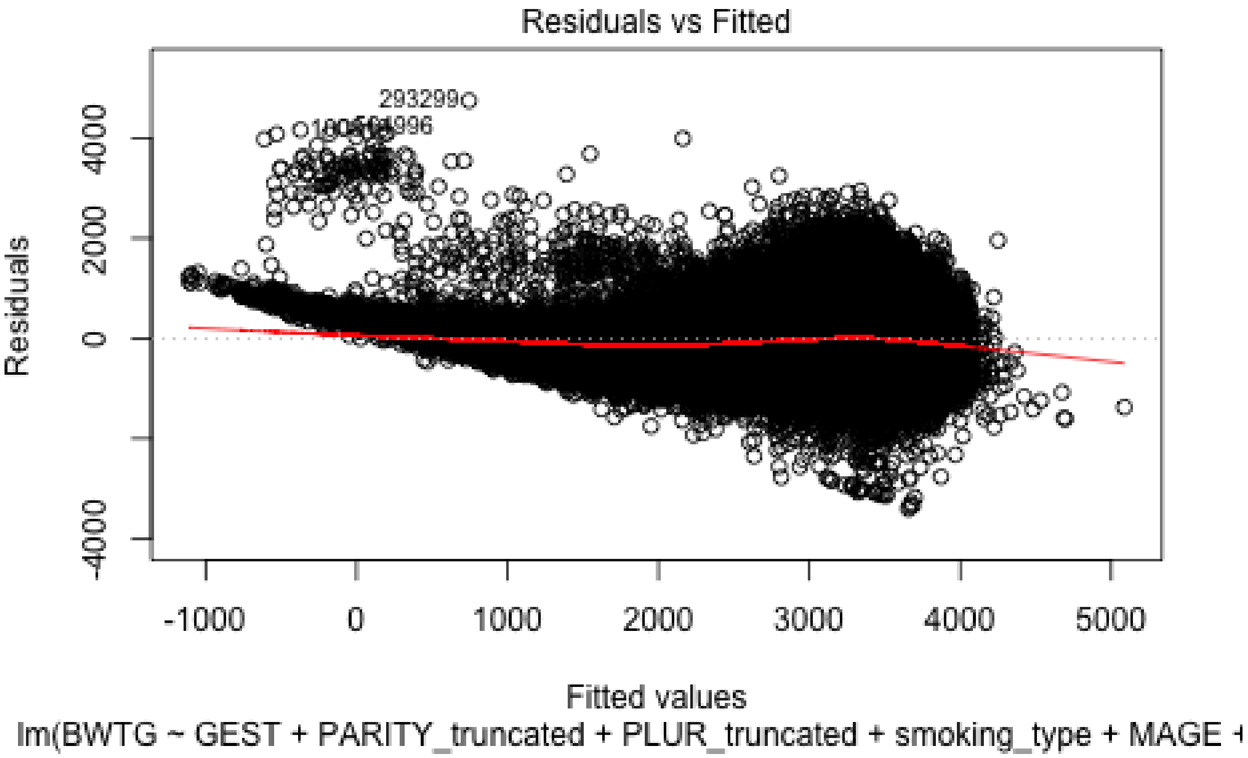
Build Model

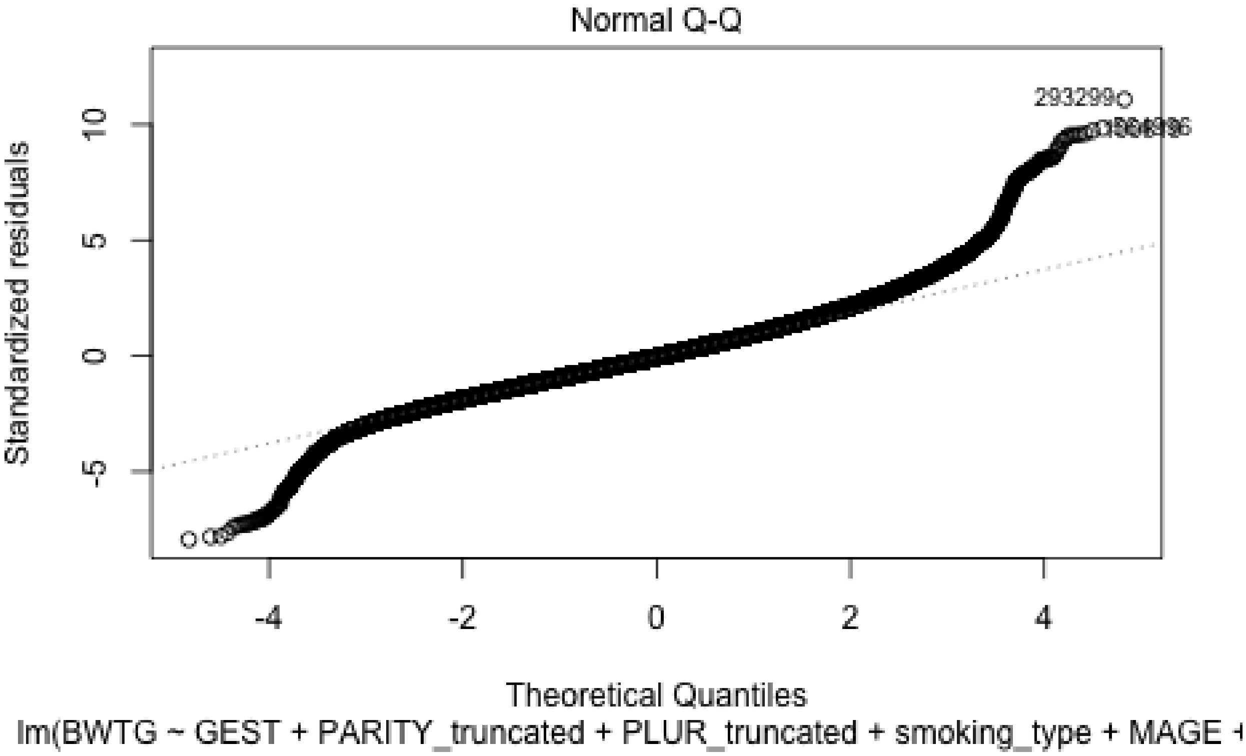
```
births_excl = na.omit(births)
births_excl = births_excl[which(births_excl$GEST < 80), ]
births_excl = births_excl %>%
  mutate(GEST2 = GEST^2, GEST3 = GEST ^ 3, GEST4 = GEST^4)
model1 = lm(data = births_excl, BWTG ~ GEST + PARITY_truncated + PLUR_truncated + smoking_type + MAGE +
summary(model1)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

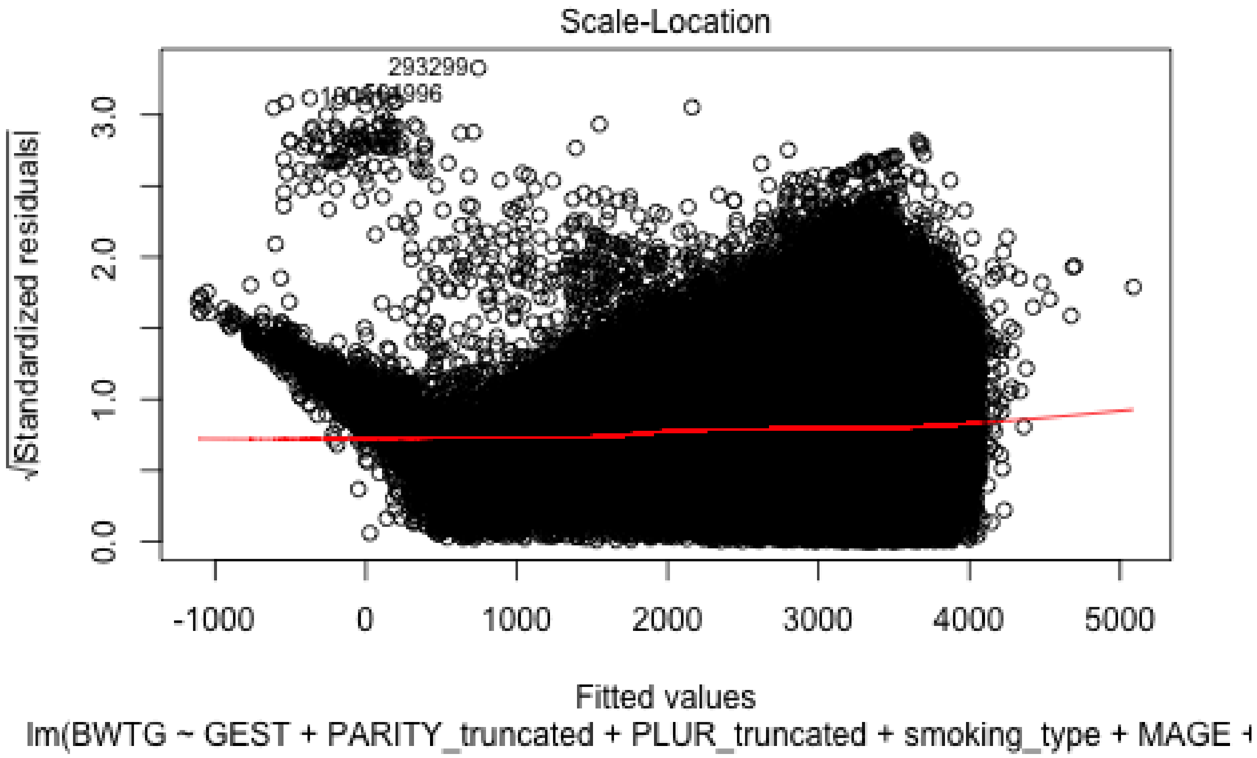
```
##
## Call:
## lm(formula = BWTG ~ GEST + PARITY_truncated + PLUR_truncated +
##     smoking_type + MAGE + MRACER + mortality, data = births_excl,
##     na.action = "na.exclude")
##
## Residuals:
```

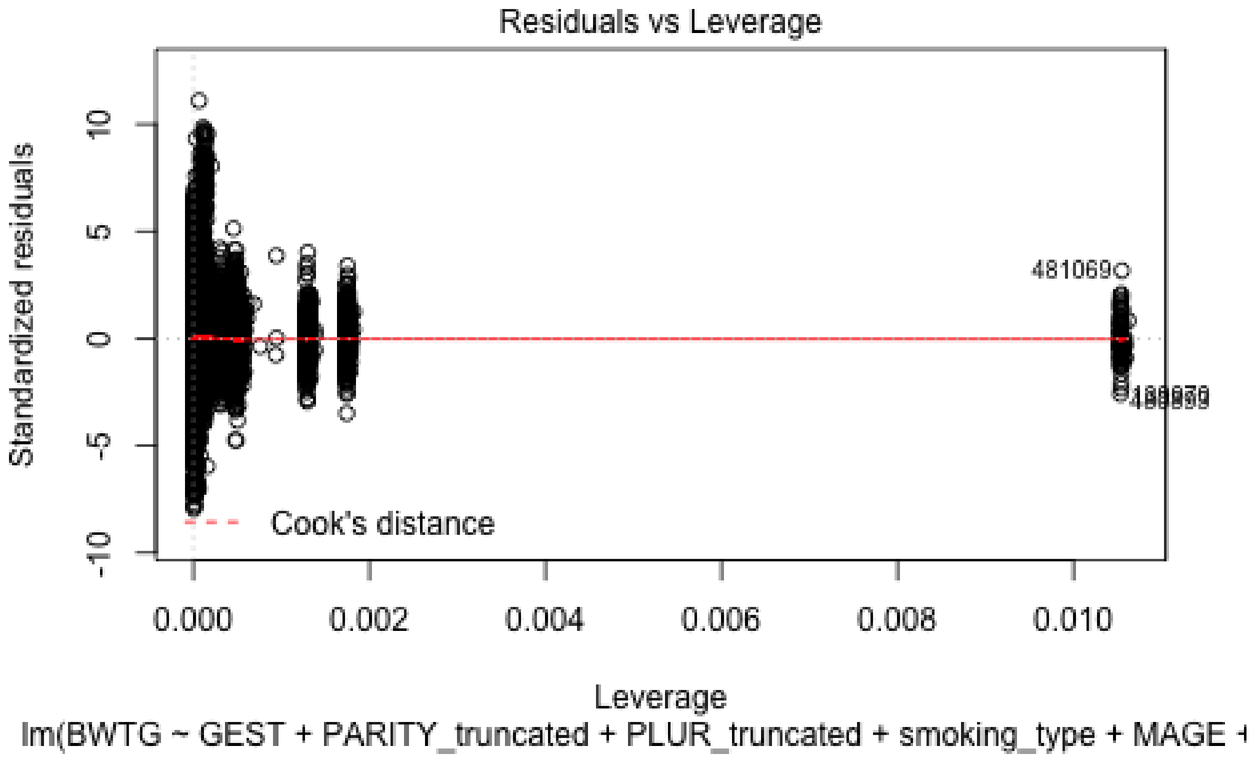
```
##     Min      1Q  Median      3Q     Max 
## -3406.6  -282.9   -19.5   260.9  4754.8 
## 
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)    
## (Intercept)                -3691.4357     9.9287 -371.794  < 2e-16 ***
## GEST                         175.2014     0.2336  749.962  < 2e-16 ***
## PARITY_truncated2             94.7007     1.3162   71.951  < 2e-16 ***
## PARITY_truncated3            113.4321     1.5379   73.756  < 2e-16 ***
## PARITY_truncated4            117.1821     1.9259   60.845  < 2e-16 ***
## PARITY_truncated5+           117.7775     1.9533   60.298  < 2e-16 ***
## PLUR_truncated2             -360.9193     2.9312 -123.128  < 2e-16 ***
## PLUR_truncated3+            -454.9182    15.3979  -29.544  < 2e-16 ***
## smoking_typebefore and during  -204.8236     1.7697 -115.737  < 2e-16 ***
## smoking_typebefore only       -11.4791     2.7026   -4.247 2.16e-05 ***
## smoking_typeduring only      -165.3562     9.1294  -18.113  < 2e-16 ***
## MAGE                           4.5889     0.0964   47.602  < 2e-16 ***
## MRACER1                       83.5128     1.6596   50.323  < 2e-16 ***
## MRACER2                      -99.3238     1.8206  -54.555  < 2e-16 ***
## MRACER3                       -4.8456     4.5757   -1.059    0.290    
## MRACER4                      -33.5469     7.6017   -4.413 1.02e-05 ***
## MRACER5                     -114.3371    17.9406   -6.373 1.85e-10 ***
## MRACER6                       24.2023    44.0275    0.550    0.583    
## MRACER7                      -14.8664     9.4582   -1.572    0.116    
## MRACER8                     -102.9048     3.2079  -32.078  < 2e-16 ***
## mortality                     23.1876     3.1019    7.475 7.72e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 428.9 on 717848 degrees of freedom
## Multiple R-squared:  0.5184, Adjusted R-squared:  0.5184 
## F-statistic: 3.864e+04 on 20 and 717848 DF,  p-value: < 2.2e-16
```

```
plot(model1)
```

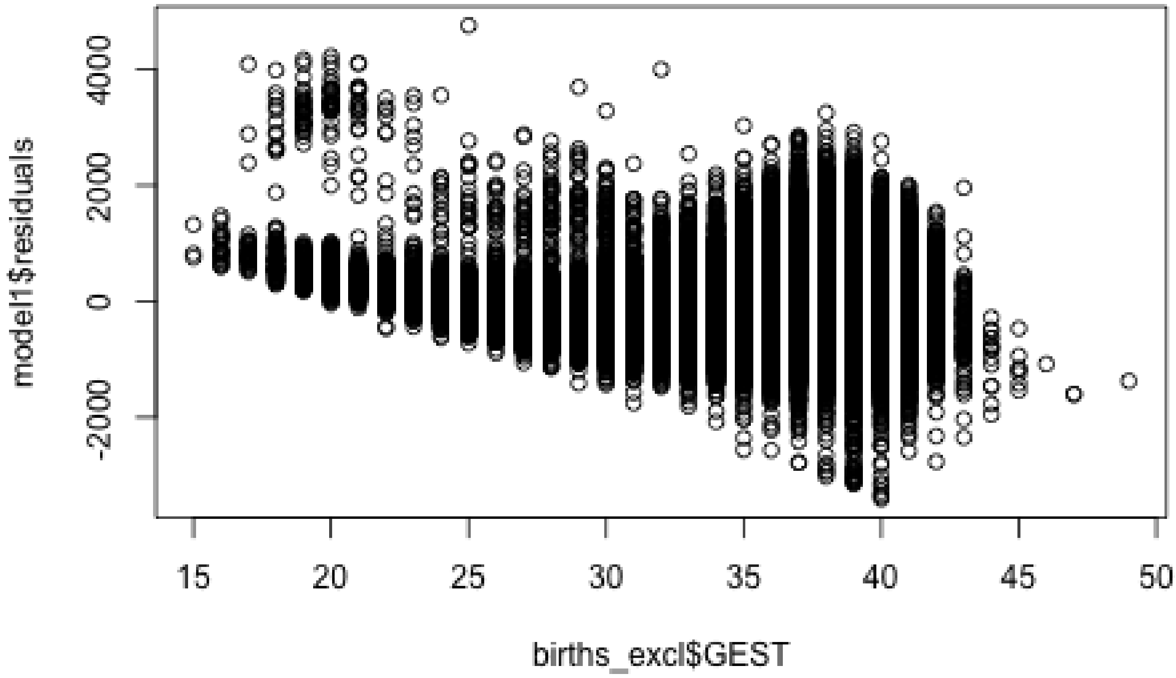

Residuals vs Fitted
Residuals
Fitted values
lm(BWTG ~ GEST + PARITY_truncated + PLUR_truncated + smoking_type + MAGE +
-4000
0
2000
4000
-1000
0
1000
2000
3000
4000
5000
293299
996


Normal Q-Q
Standardized residuals
10
5
0
-5
293299
-4
-2
0
2
4
Theoretical Quantiles
lm(BWTG ~ GEST + PARITY_truncated + PLUR_truncated + smoking_type + MAGE +

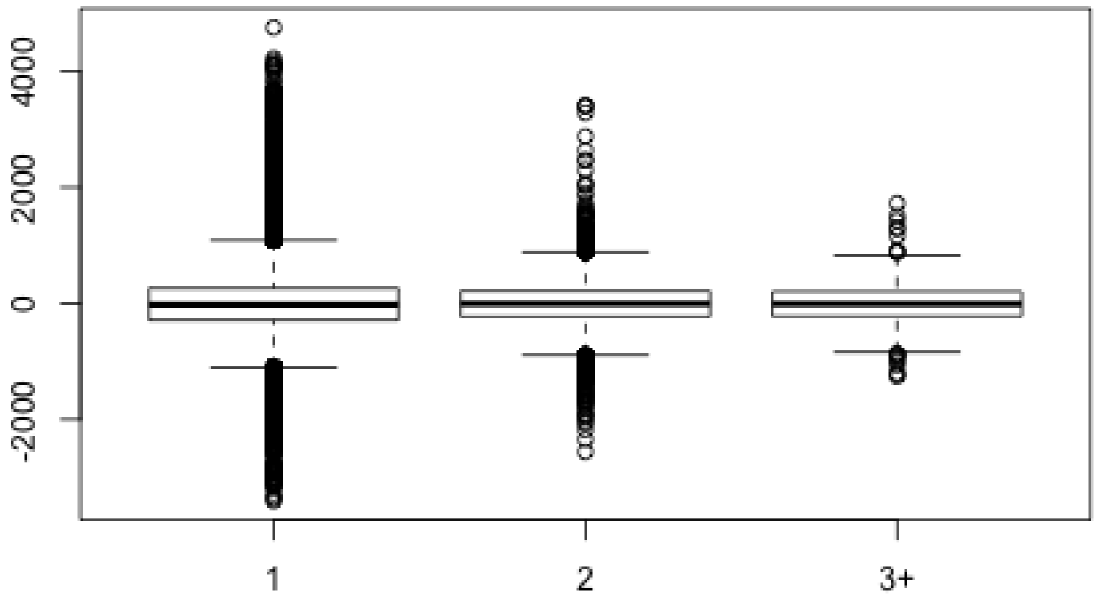

Scale-Location
293299
√|Standardized residuals|
0.0
1.0
2.0
3.0
-1000
0
1000
2000
3000
4000
5000
Fitted values
lm(BWTG ~ GEST + PARITY_truncated + PLUR_truncated + smoking_type + MAGE +



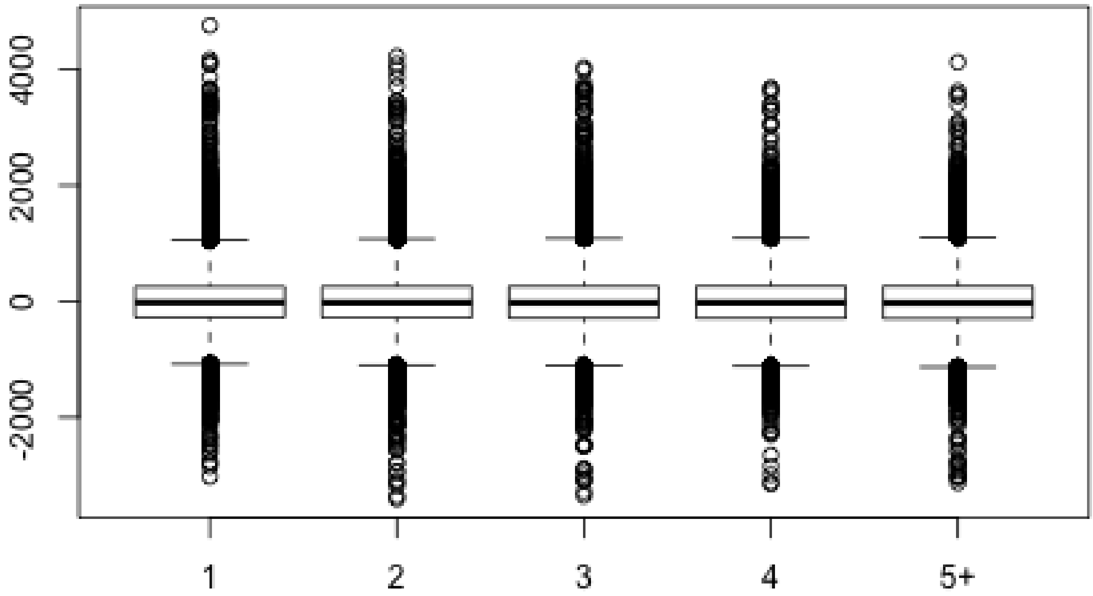
```
# plot(model1$fitted.values, model1$residuals)
plot(births_excl$GEST, model1$residuals)
```



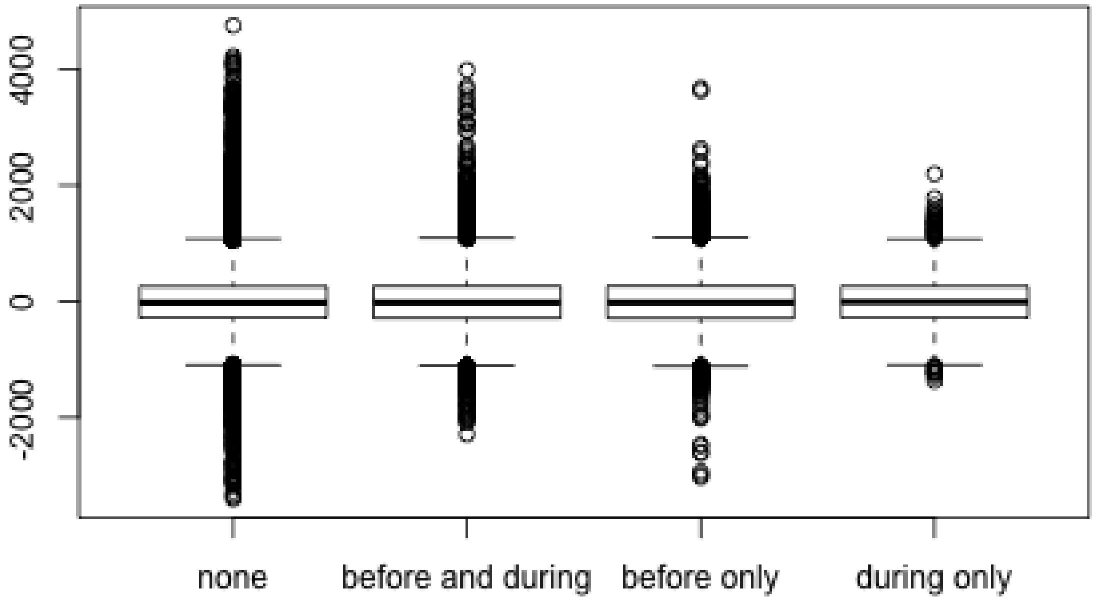
```
plot(births_excl$PLUR_truncated, model1$residuals)
```



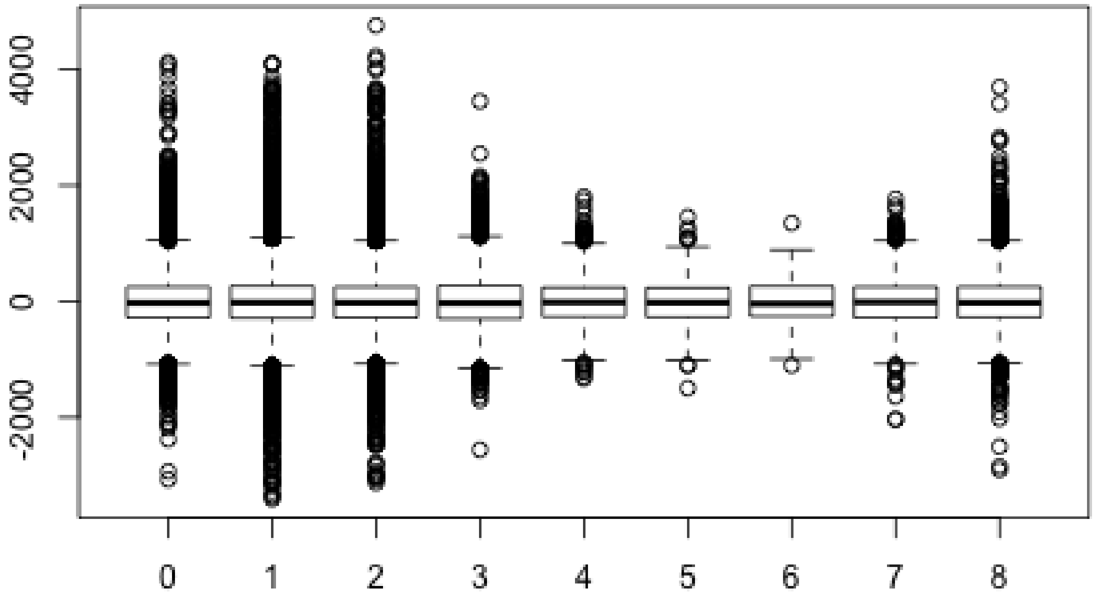
```r
plot(births_excl$PARITY_truncated, model1$residuals)
```



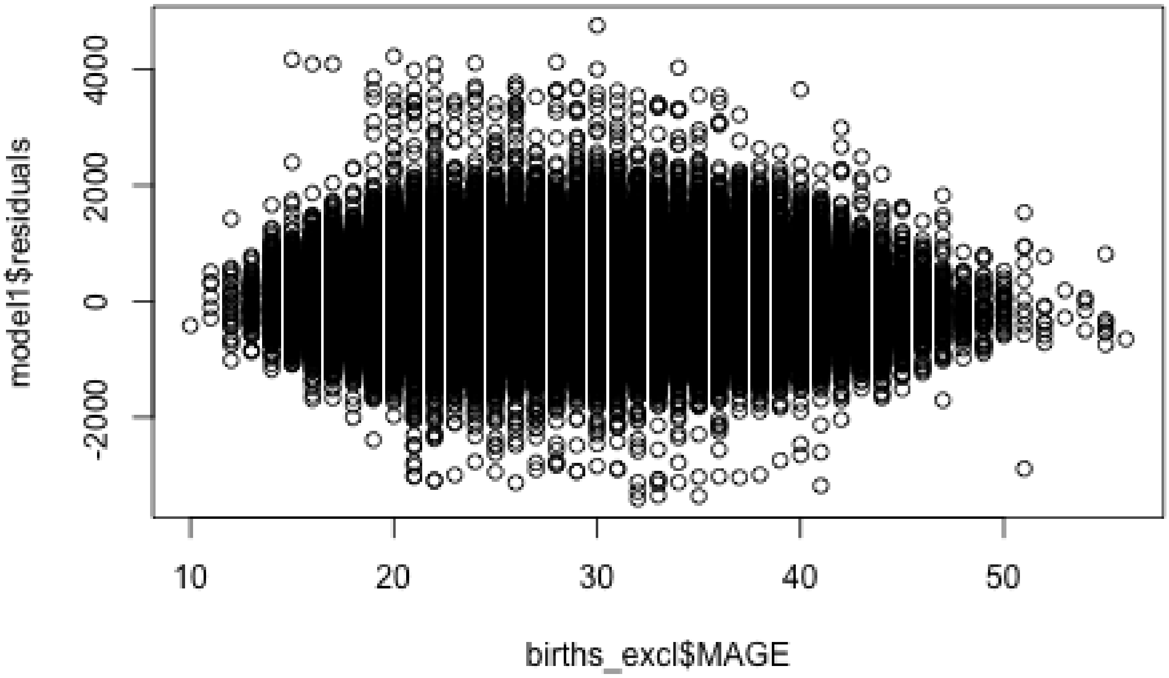
```r
plot(births_excl$smoking_type, model1$residuals)
```



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



```r
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.

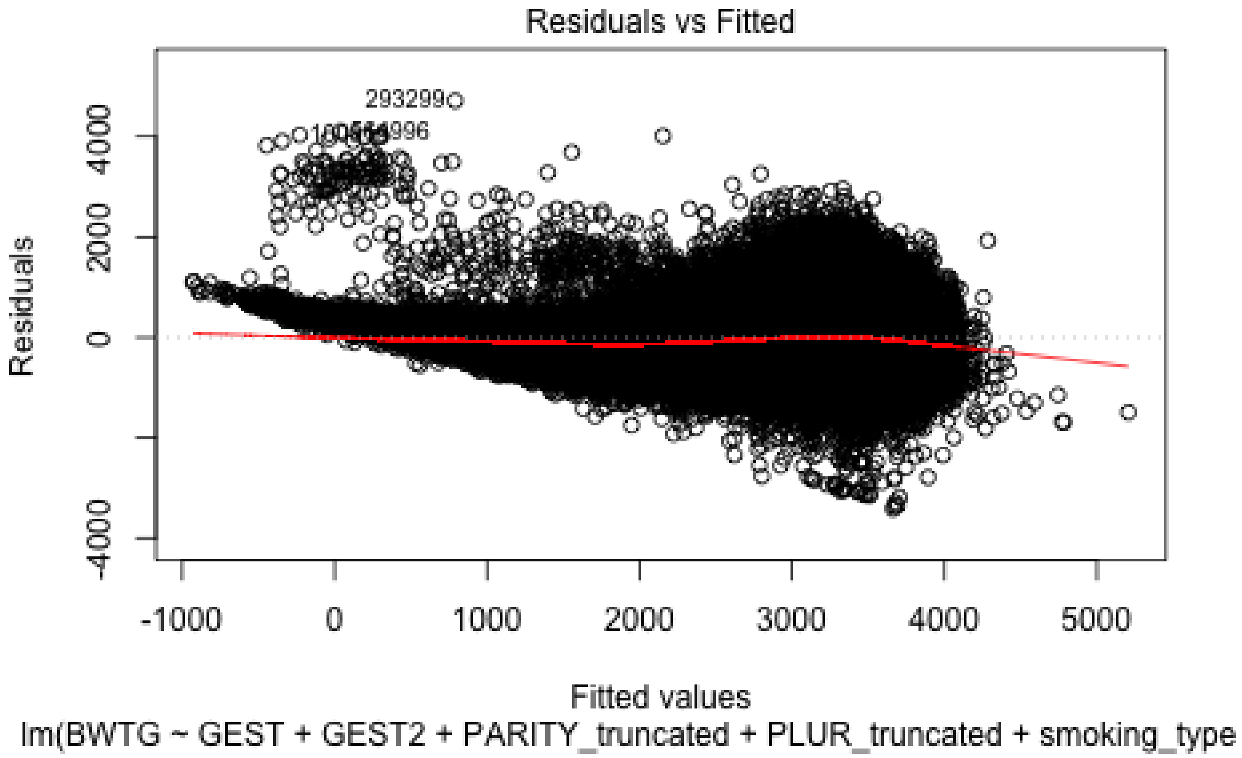
The residual graph for Mother's age is fairly random, with residuals getting a bit smaller near the beginning and end (<20 years old and >45 years old).

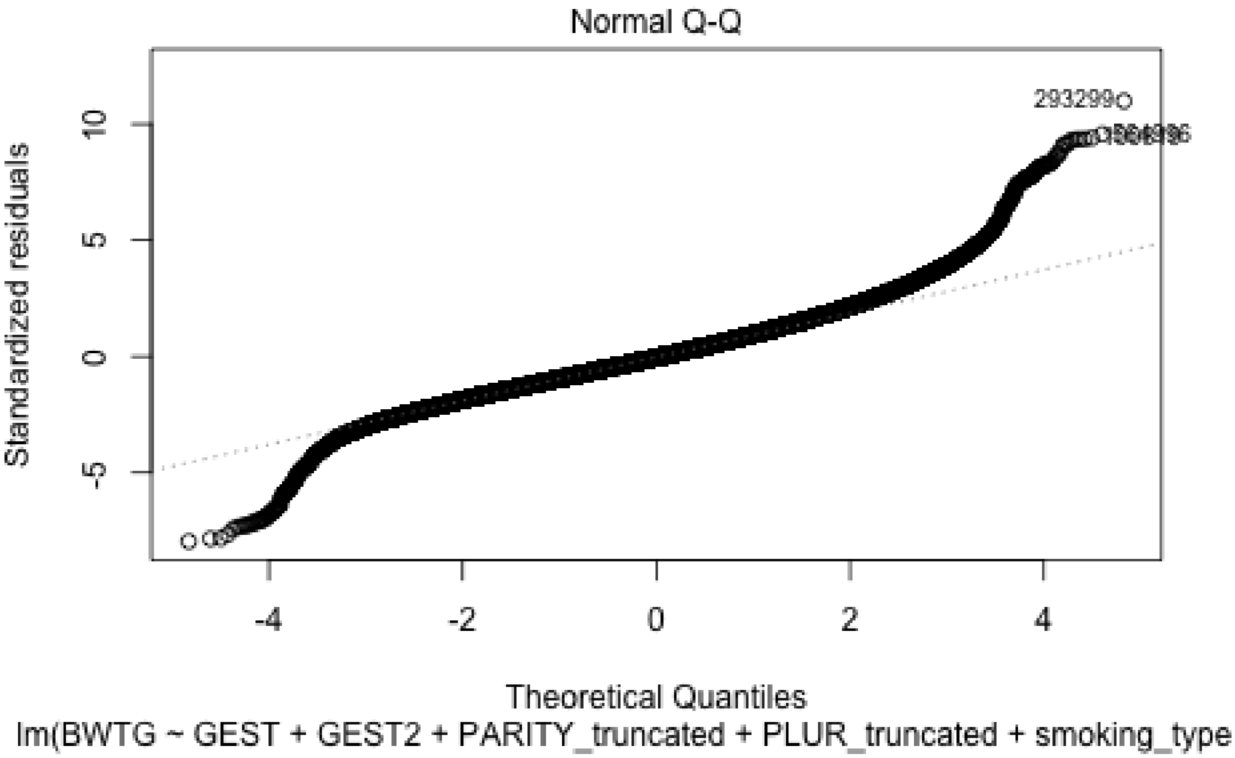
```r
model2 = lm(data = births_excl, BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated + smoking_type
summary(model2)
```

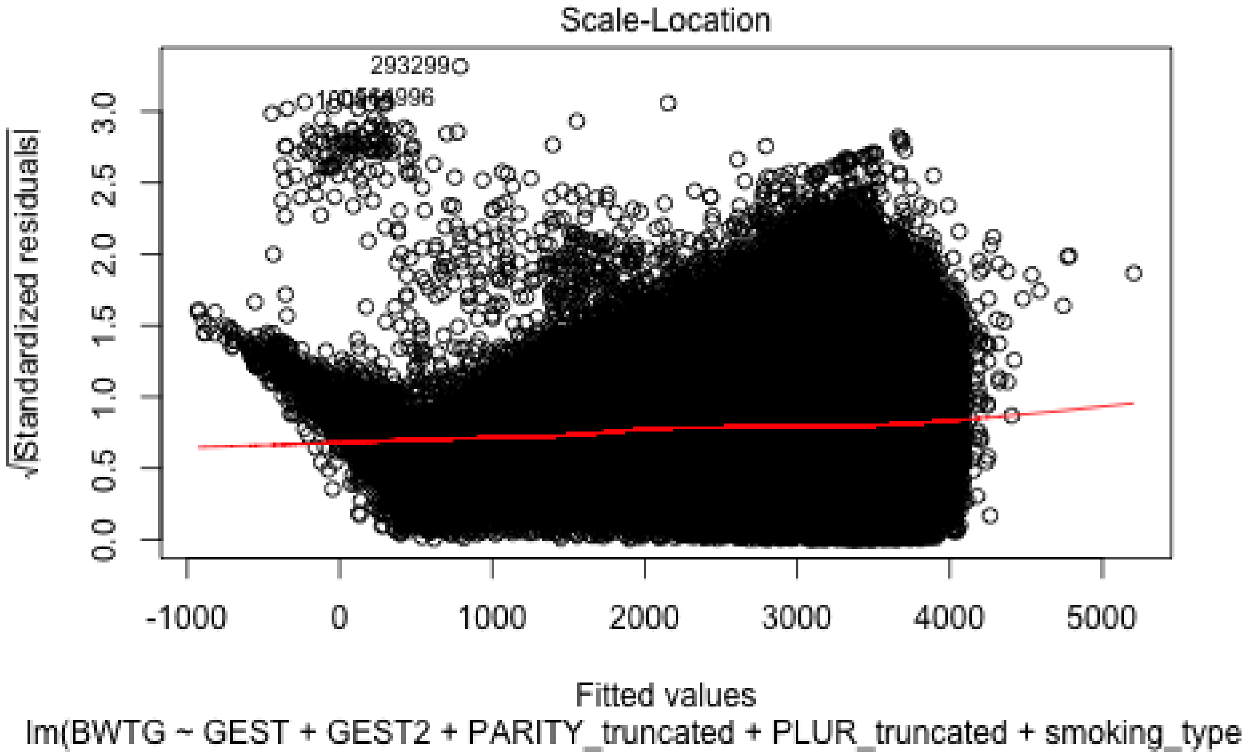
```
##
```

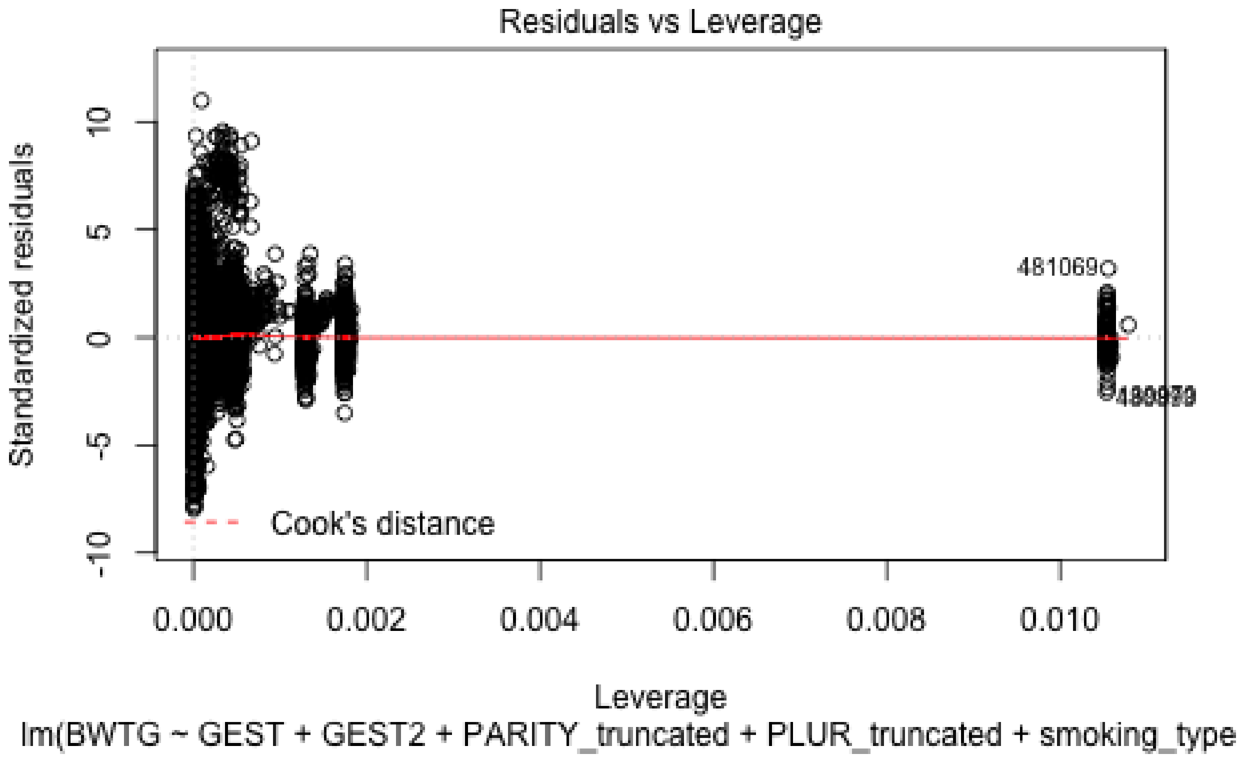
```
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated +
##     smoking_type + MAGE + MRACER + mortality, data = births_excl,
##     na.action = "na.exclude")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3412.1  -282.9   -19.5   260.8  4709.9
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 -2.938e+03  4.046e+01  -72.608  < 2e-16 ***
## GEST                         1.307e+02  2.327e+00   56.181  < 2e-16 ***
## GEST2                        6.427e-01  3.346e-02   19.208  < 2e-16 ***
## PARITY_truncated2            9.593e+01  1.317e+00   72.817  < 2e-16 ***
## PARITY_truncated3            1.149e+02  1.540e+00   74.659  < 2e-16 ***
## PARITY_truncated4            1.190e+02  1.928e+00   61.738  < 2e-16 ***
## PARITY_truncated5+           1.198e+02  1.956e+00   61.256  < 2e-16 ***
## PLUR_truncated2             -3.558e+02  2.943e+00 -120.901  < 2e-16 ***
## PLUR_truncated3+            -4.566e+02  1.539e+01  -29.658  < 2e-16 ***
## smoking_typebefore and during -2.041e+02  1.770e+00 -115.320  < 2e-16 ***
## smoking_typebefore only       -1.152e+01  2.702e+00   -4.264 2.00e-05 ***
## smoking_typeduring only       -1.651e+02  9.127e+00  -18.087  < 2e-16 ***
## MAGE                         4.611e+00  9.638e-02   47.841  < 2e-16 ***
## MRACER1                      8.373e+01  1.659e+00   50.463  < 2e-16 ***
## MRACER2                     -9.902e+01  1.820e+00  -54.402  < 2e-16 ***
## MRACER3                     -5.140e+00  4.575e+00   -1.124    0.261
## MRACER4                     -3.340e+01  7.600e+00   -4.395 1.11e-05 ***
## MRACER5                     -1.135e+02  1.794e+01   -6.329 2.47e-10 ***
## MRACER6                      2.394e+01  4.402e+01    0.544    0.586
## MRACER7                     -1.327e+01  9.456e+00   -1.403    0.161
## MRACER8                     -1.020e+02  3.207e+00  -31.797  < 2e-16 ***
## mortality                    2.390e+01  3.101e+00    7.707 1.29e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 428.8 on 717847 degrees of freedom
## Multiple R-squared:  0.5187, Adjusted R-squared:  0.5186
## F-statistic: 3.683e+04 on 21 and 717847 DF,  p-value: < 2.2e-16
```

```
plot(model2)
```

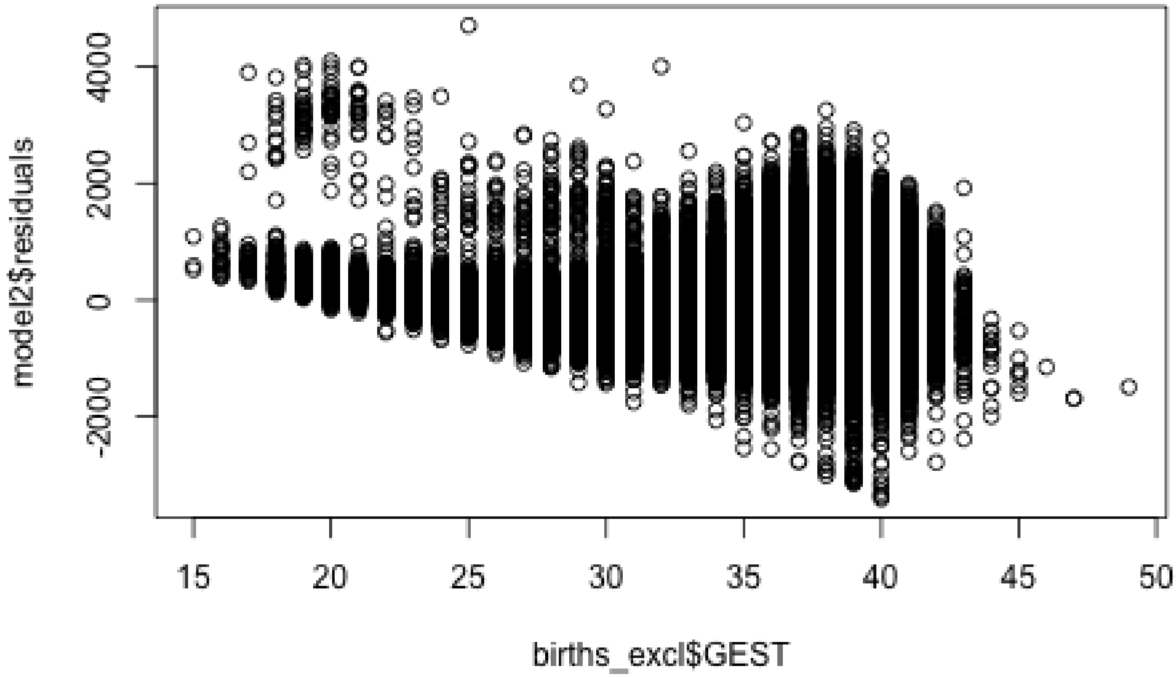

Residuals vs Fitted
293299
Residuals
4000
2000
0
-4000
-1000
0
1000
2000
3000
4000
5000
Fitted values
lm(BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated + smoking_type


Normal Q-Q
293299
Standardized residuals
10
5
0
-5
-4
-2
0
2
4
Theoretical Quantiles
lm(BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated + smoking_type

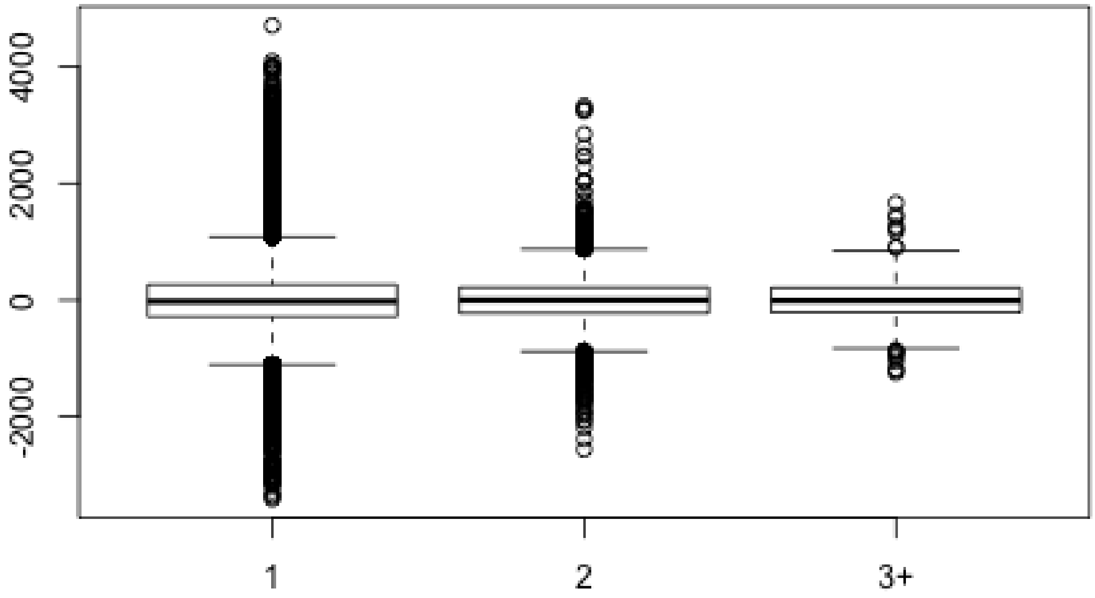

Scale-Location
293299
180514996
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-1000
0
1000
2000
3000
4000
5000
Fitted values
lm(BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated + smoking_type



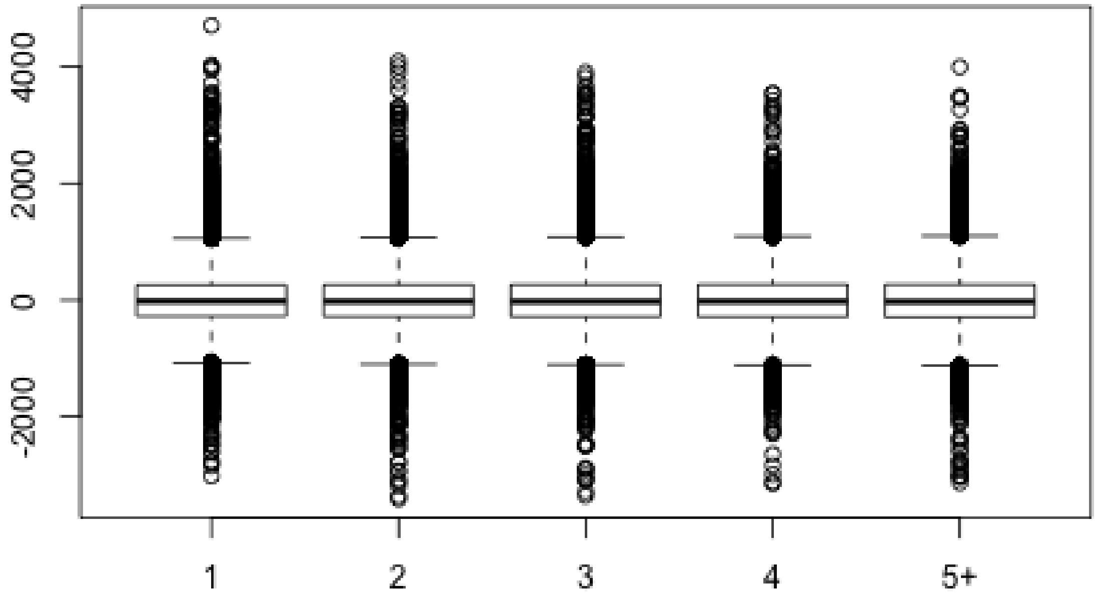
```r
# plot(model2$fitted.values, model2$residuals)
plot(births_excl$GEST, model2$residuals)
```



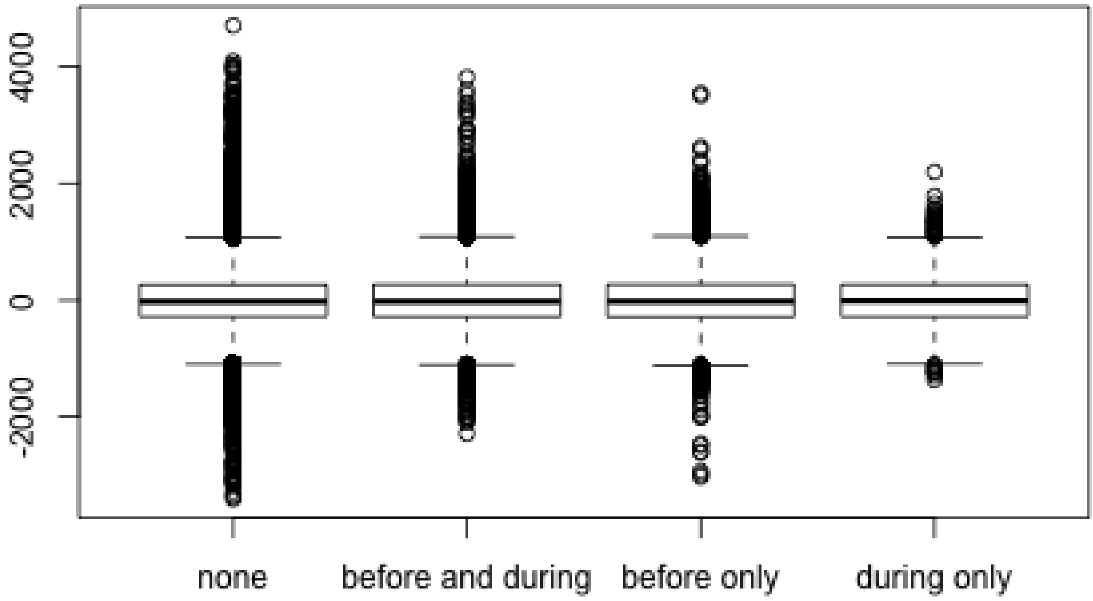
```r
plot(births_excl$PLUR_truncated, model2$residuals)
```



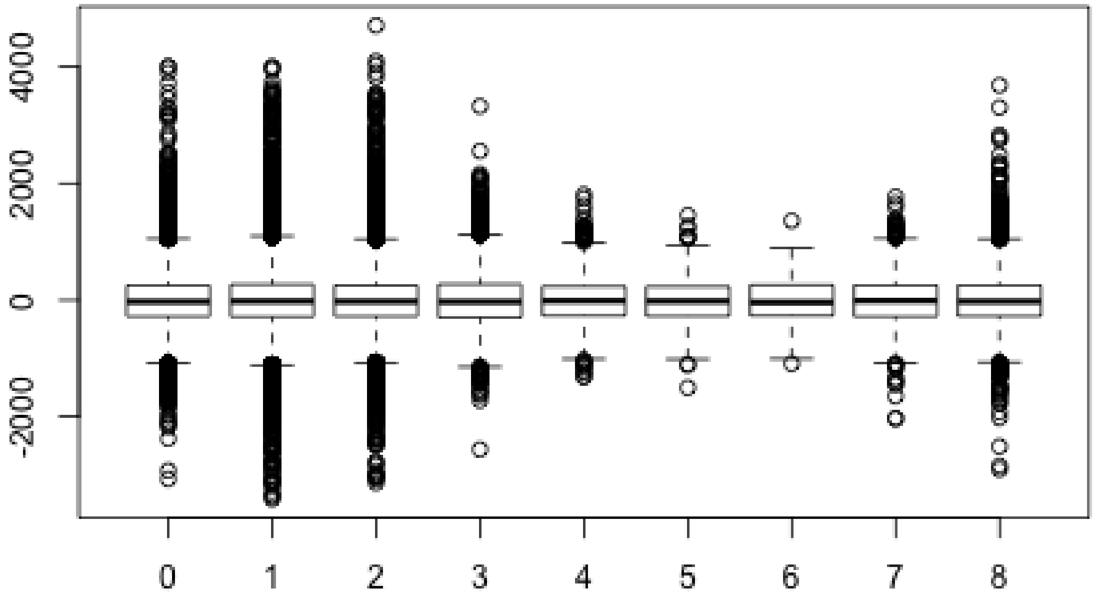
```
plot(births_excl$PARITY_truncated, model2$residuals)
```



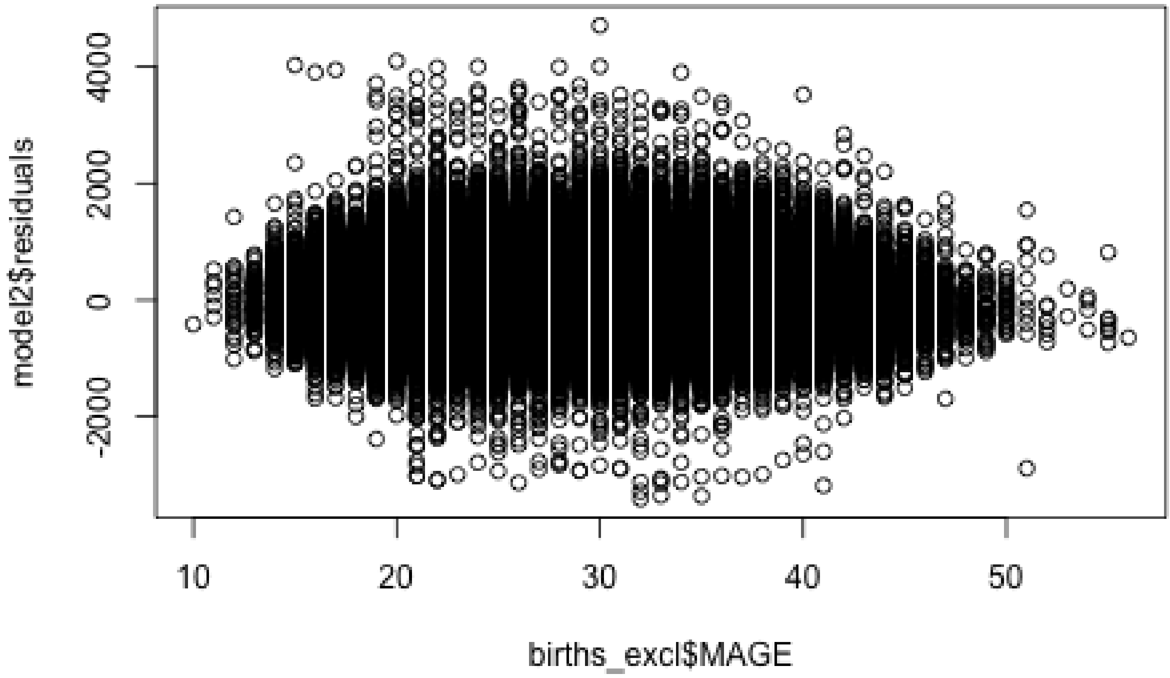
```r
plot(births_excl$smoking_type, model2$residuals)
```



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



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

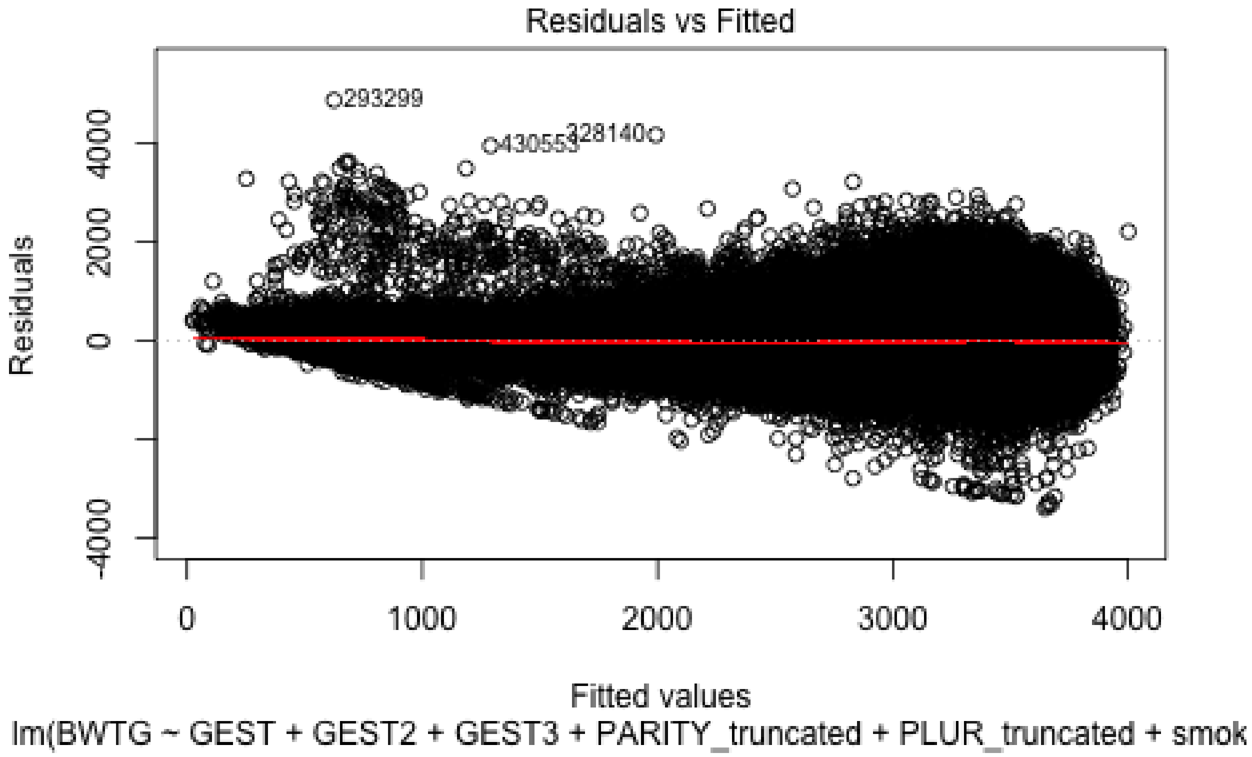


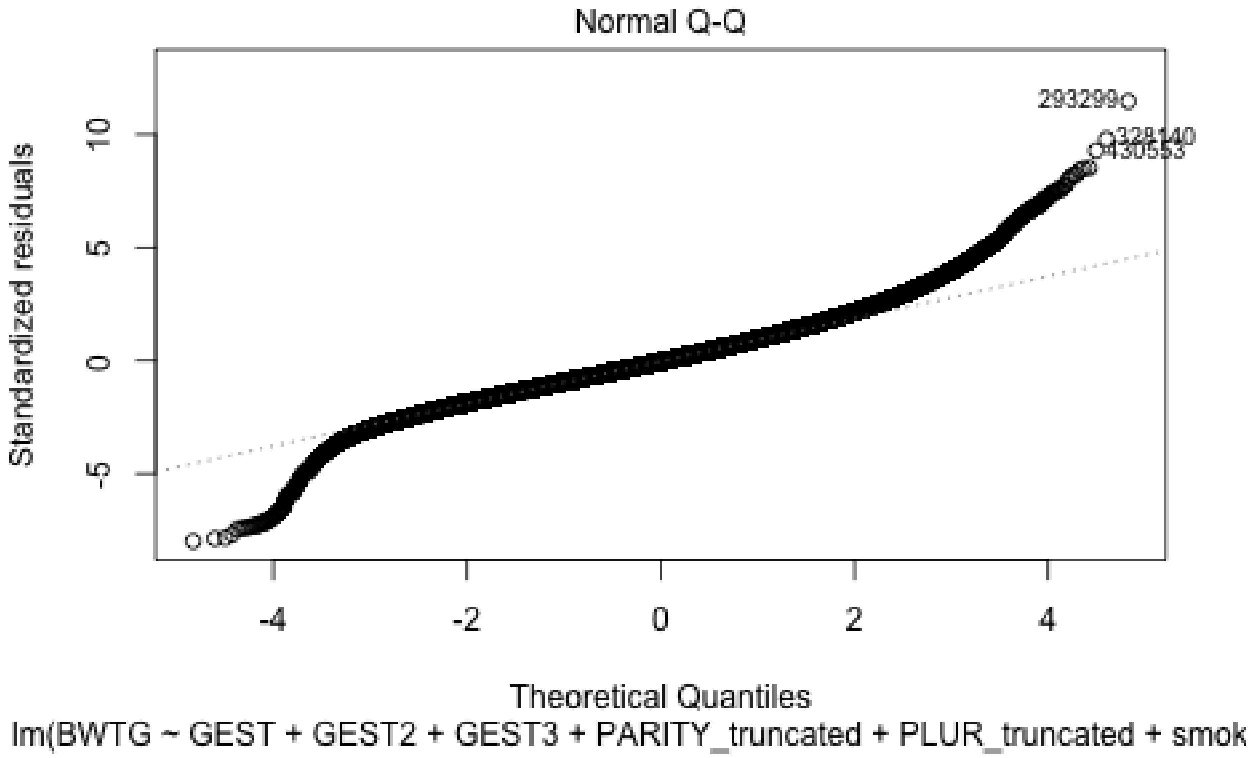
The new model still displays the original downwards trend in the residual vs gestational period graph. Perhaps another transformation on GEST would be helpful – a cubic term can be added. The other residuals plots also retain their trends from model 1.

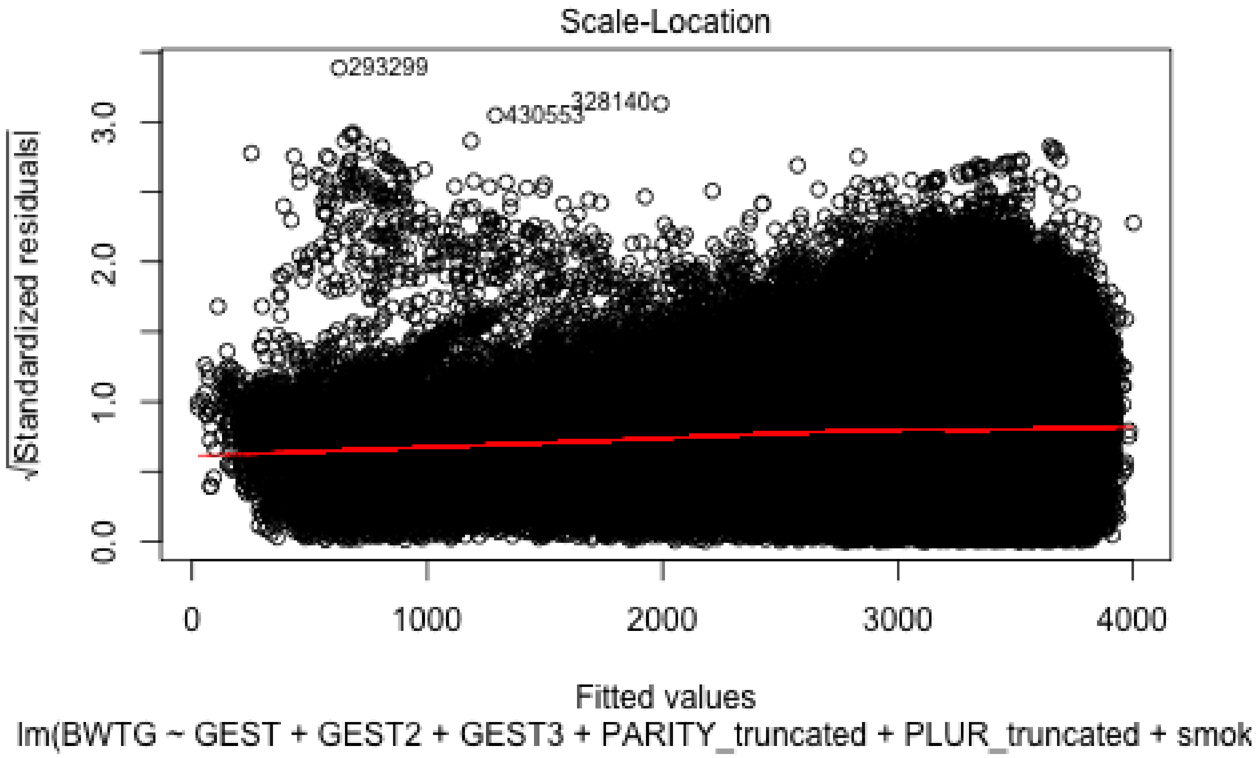
```
model3 = lm(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated + PLUR_truncated + smoki
summary(model3)
```

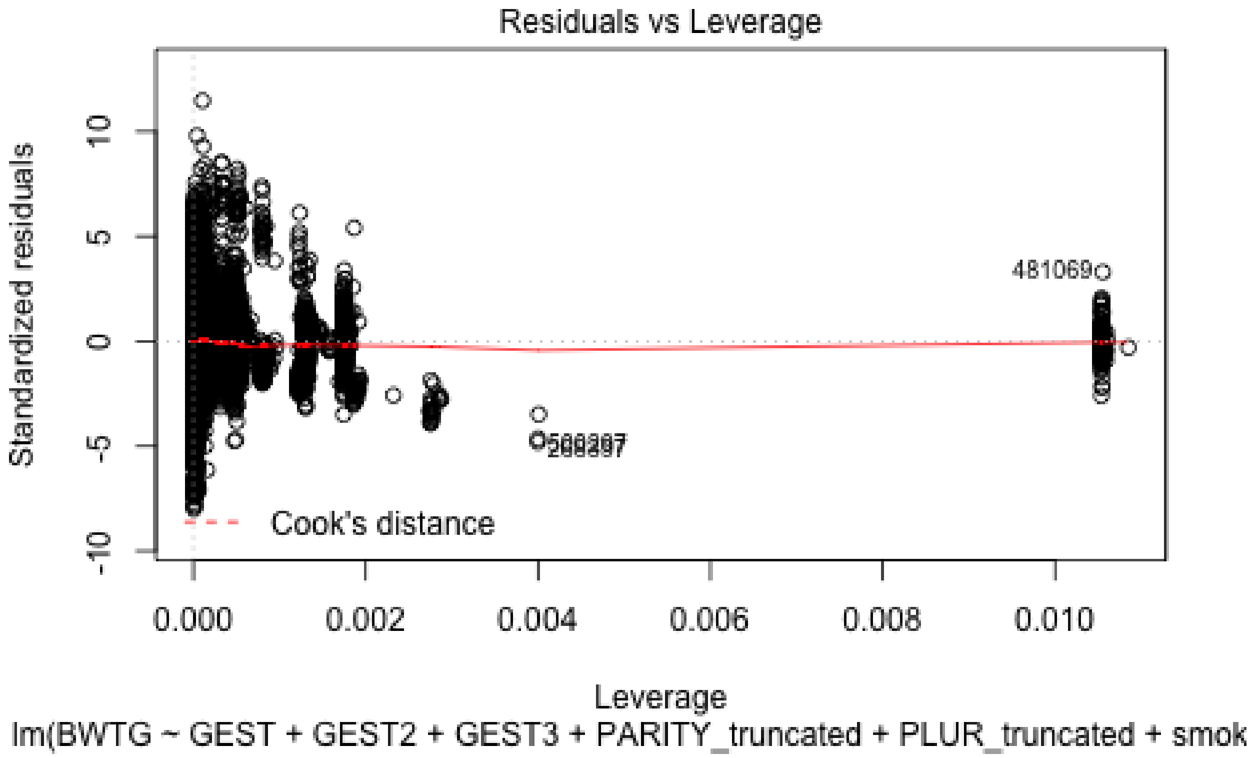
```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3396.9  -280.5   -18.7   258.0  4875.2
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                   1.701e+04  1.866e+02   91.157  < 2e-16 ***
## GEST                         -1.785e+03  1.765e+01 -101.116  < 2e-16 ***
## GEST2                         6.013e+01  5.445e-01  110.439  < 2e-16 ***
## GEST3                        -6.018e-01  5.498e-03 -109.462  < 2e-16 ***
## PARITY_truncated2             8.815e+01  1.308e+00   67.366  < 2e-16 ***
## PARITY_truncated3             1.055e+02  1.529e+00   68.985  < 2e-16 ***
## PARITY_truncated4             1.100e+02  1.914e+00   57.459  < 2e-16 ***
## PARITY_truncated5+            1.118e+02  1.941e+00   57.630  < 2e-16 ***
## PLUR_truncated2              -3.303e+02  2.928e+00 -112.810  < 2e-16 ***
## PLUR_truncated3+             -3.485e+02  1.530e+01  -22.780  < 2e-16 ***
## smoking_typebefore and during -2.033e+02  1.755e+00 -115.846  < 2e-16 ***
## smoking_typebefore only       -1.130e+01  2.680e+00   -4.218 2.46e-05 ***
## smoking_typeduring only       -1.651e+02  9.052e+00  -18.240  < 2e-16 ***
## MAGE                           4.592e+00  9.559e-02   48.040  < 2e-16 ***
## MRACER1                        8.254e+01  1.646e+00   50.158  < 2e-16 ***
## MRACER2                       -9.966e+01  1.805e+00  -55.206  < 2e-16 ***
## MRACER3                       -2.102e+00  4.537e+00   -0.463   0.6431
## MRACER4                       -3.828e+01  7.537e+00   -5.079 3.80e-07 ***
## MRACER5                       -1.184e+02  1.779e+01   -6.658 2.78e-11 ***
## MRACER6                        2.413e+01  4.365e+01    0.553   0.5805
## MRACER7                       -1.951e+01  9.378e+00   -2.080   0.0375 *
## MRACER8                       -1.070e+02  3.181e+00  -33.621  < 2e-16 ***
## mortality                      1.952e+01  3.076e+00    6.345 2.22e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 425.2 on 717846 degrees of freedom
## Multiple R-squared:  0.5266, Adjusted R-squared:  0.5265
## F-statistic: 3.629e+04 on 22 and 717846 DF,  p-value: < 2.2e-16
```

```r
plot(model3)
```

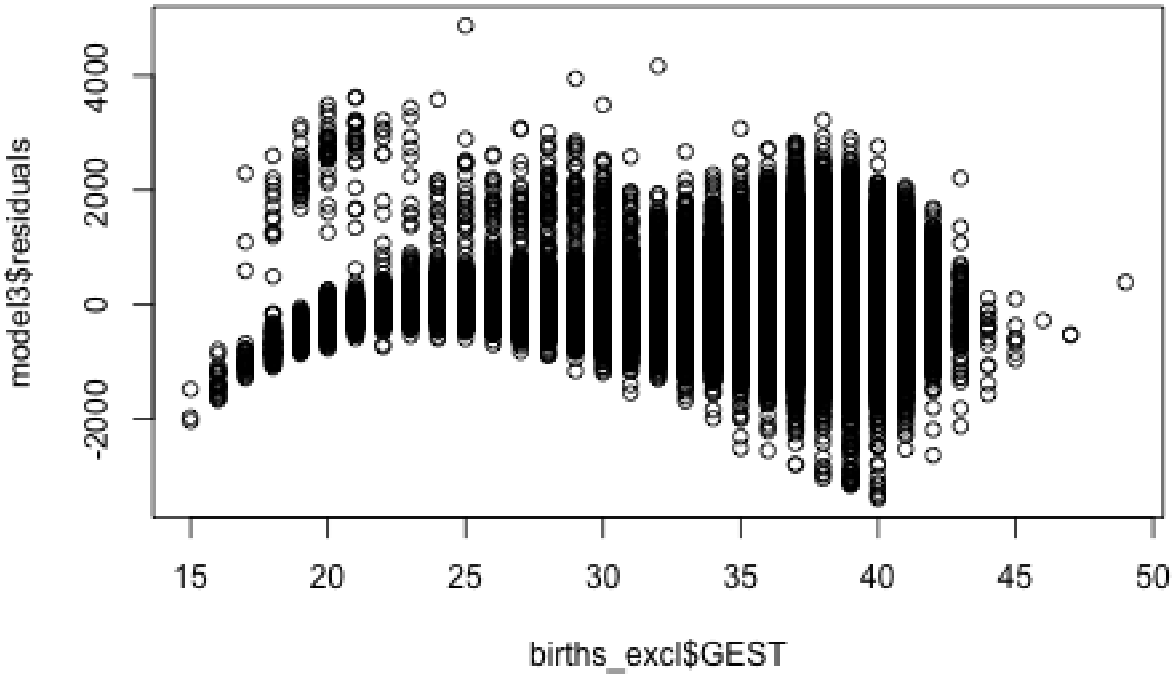



Normal Q-Q
Standardized residuals
10
5
0
-5
-4
-2
0
2
4
293299
328140
430553
Theoretical Quantiles
lm(BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated + PLUR_truncated + smok

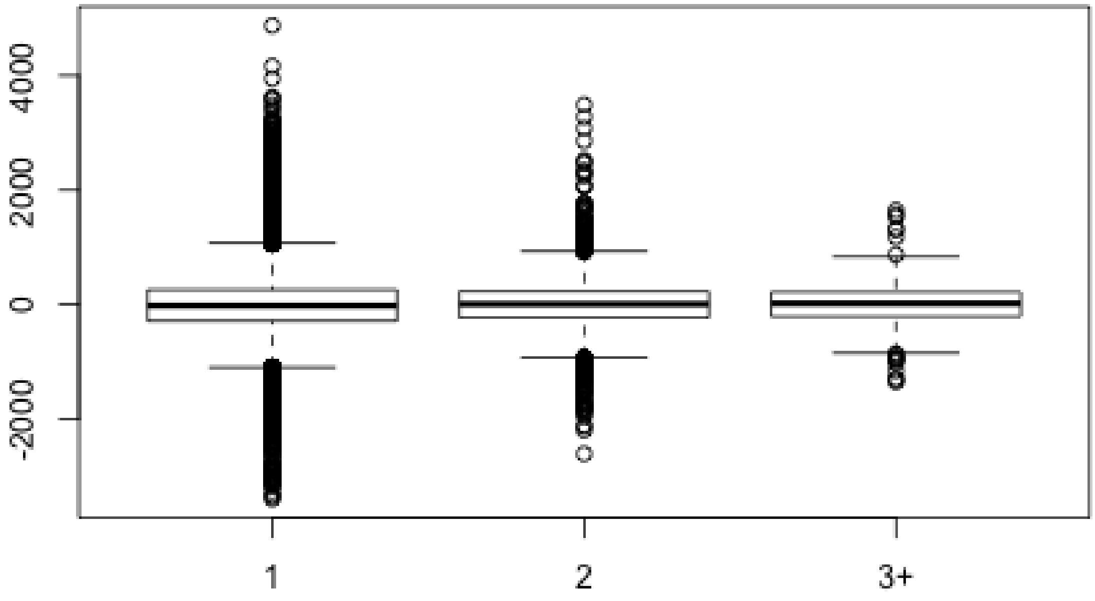

Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
3.0
0
1000
2000
3000
4000
Fitted values
lm(BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated + PLUR_truncated + smok
293299
430553
328140



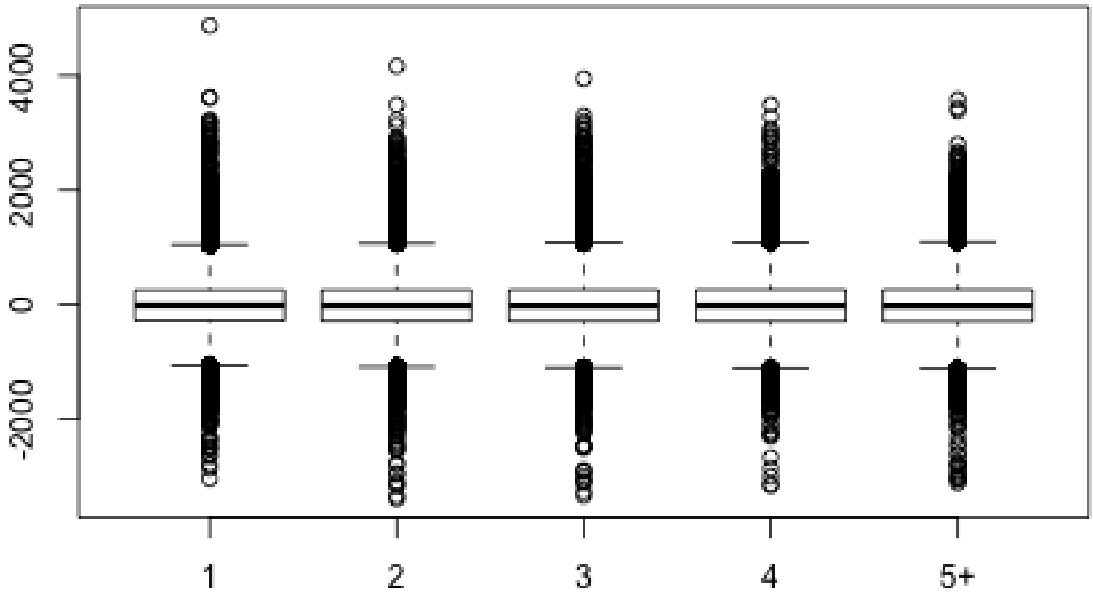
```r
# plot(model3$fitted.values, model3$residuals)
plot(births_excl$GEST, model3$residuals)
```



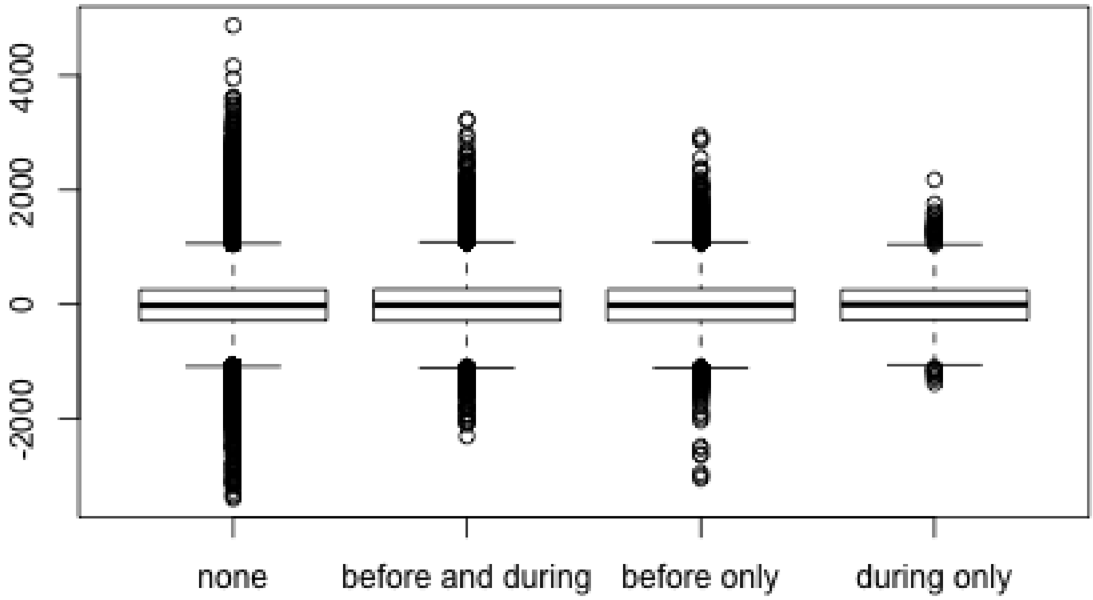
```
plot(births_excl$PLUR_truncated, model3$residuals)
```



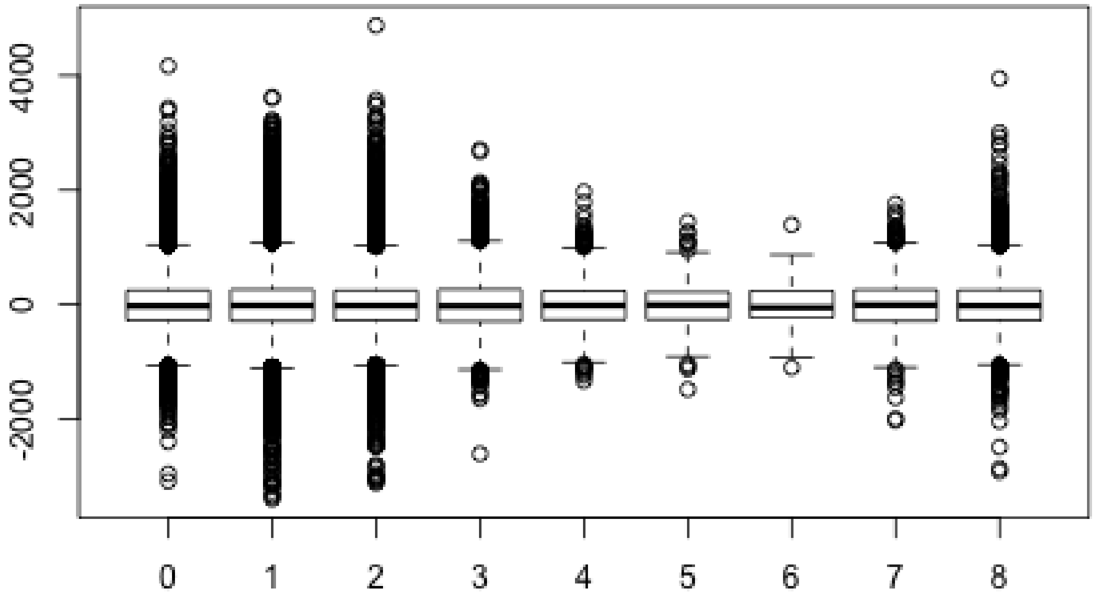
```
plot(births_excl$PARITY_truncated, model3$residuals)
```



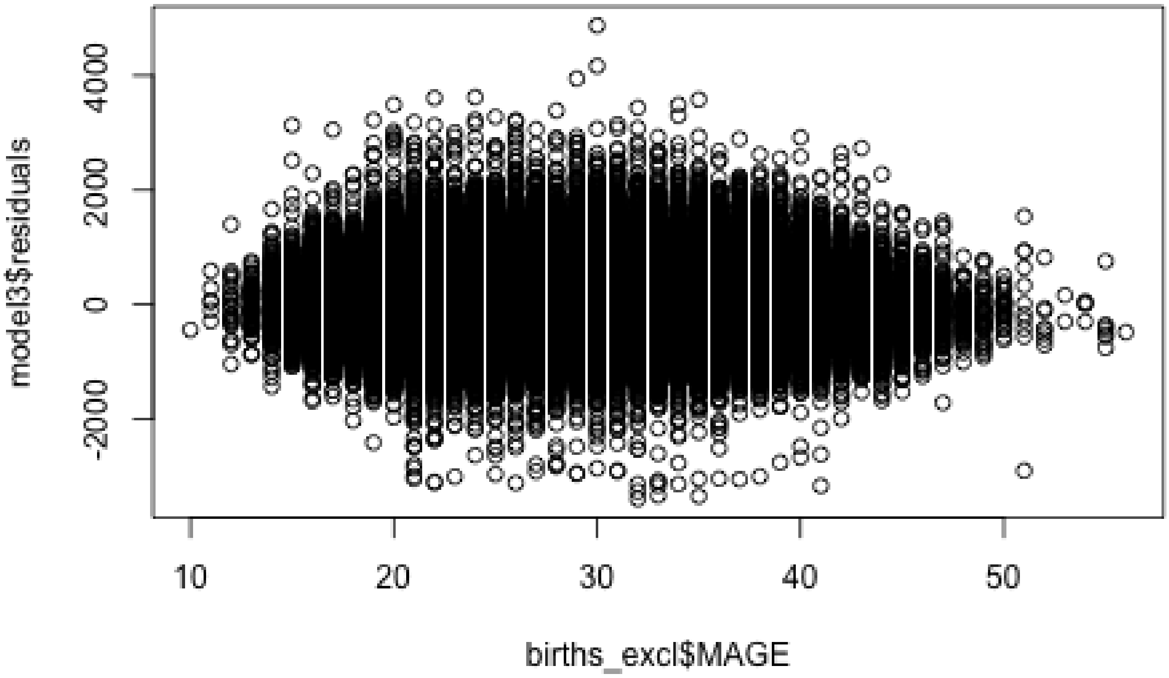
```r
plot(births_excl$smoking_type, model3$residuals)
```



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



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

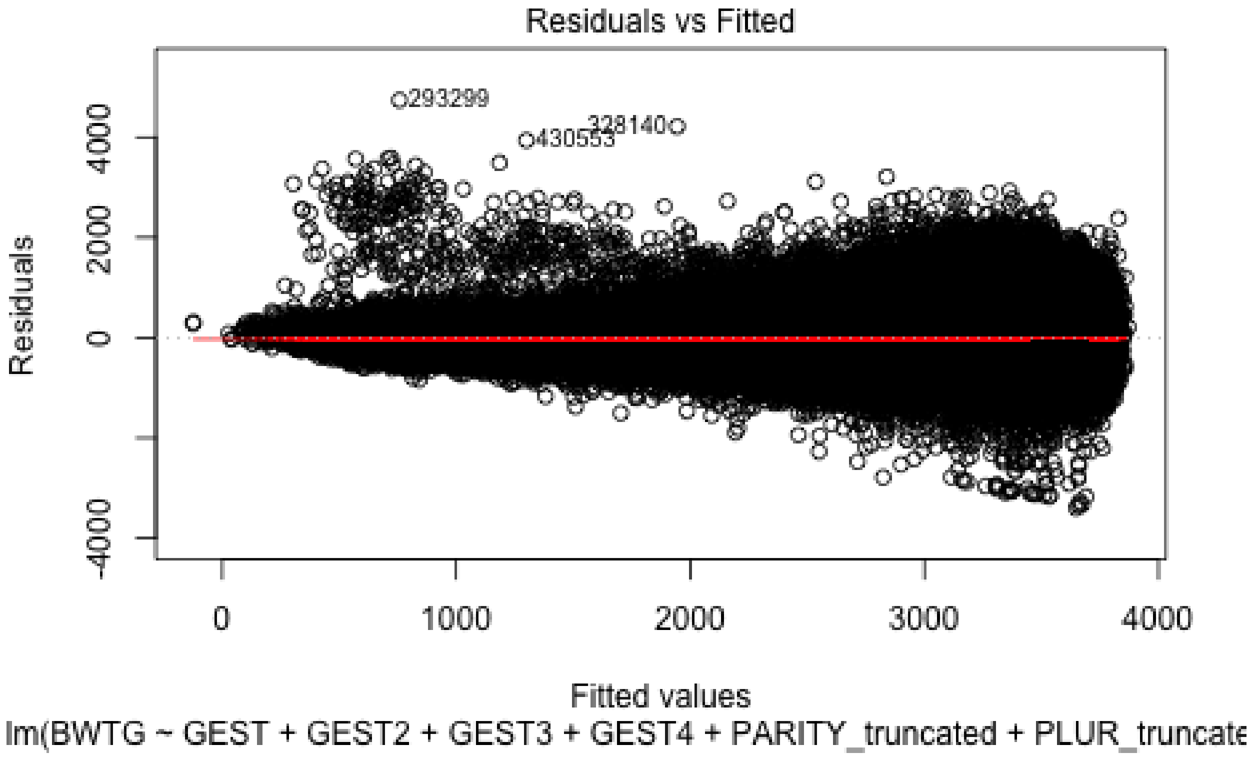


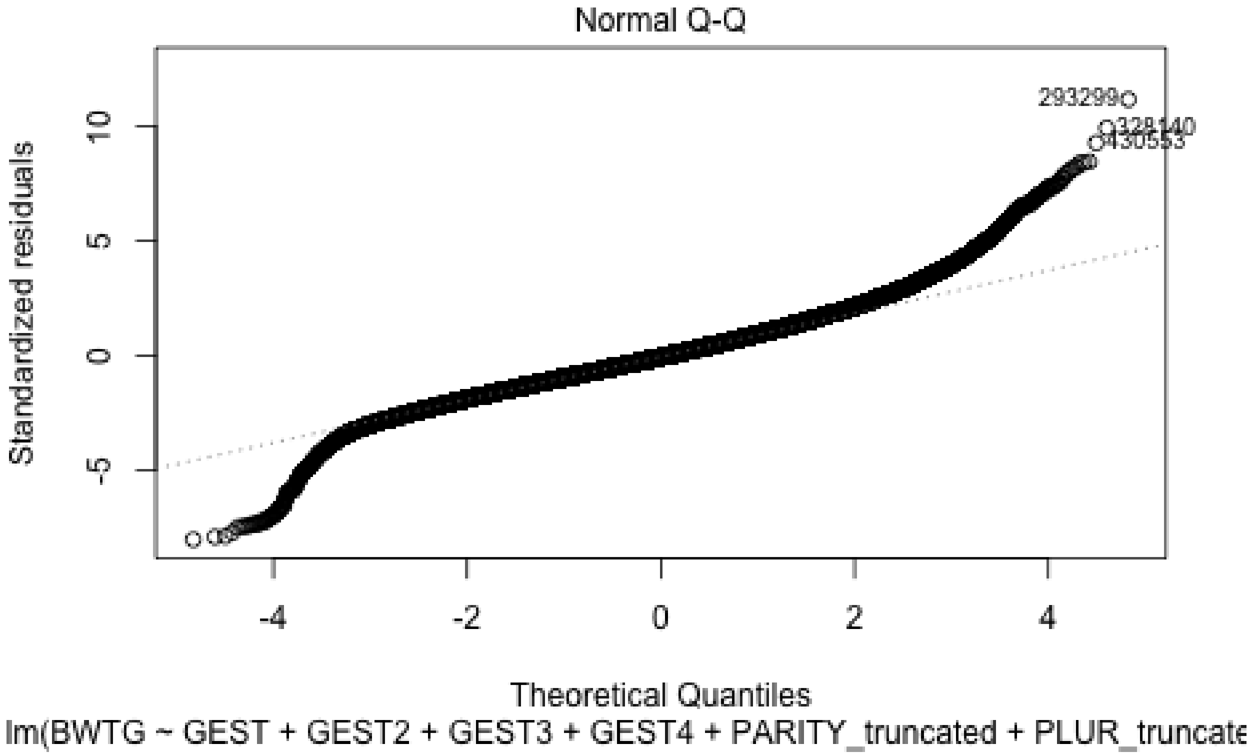
The residual vs gestational period shows a much more random pattern than before. It is worth investigating if adding a quartic term would help. The other residuals plots also retain their trends from model 1.

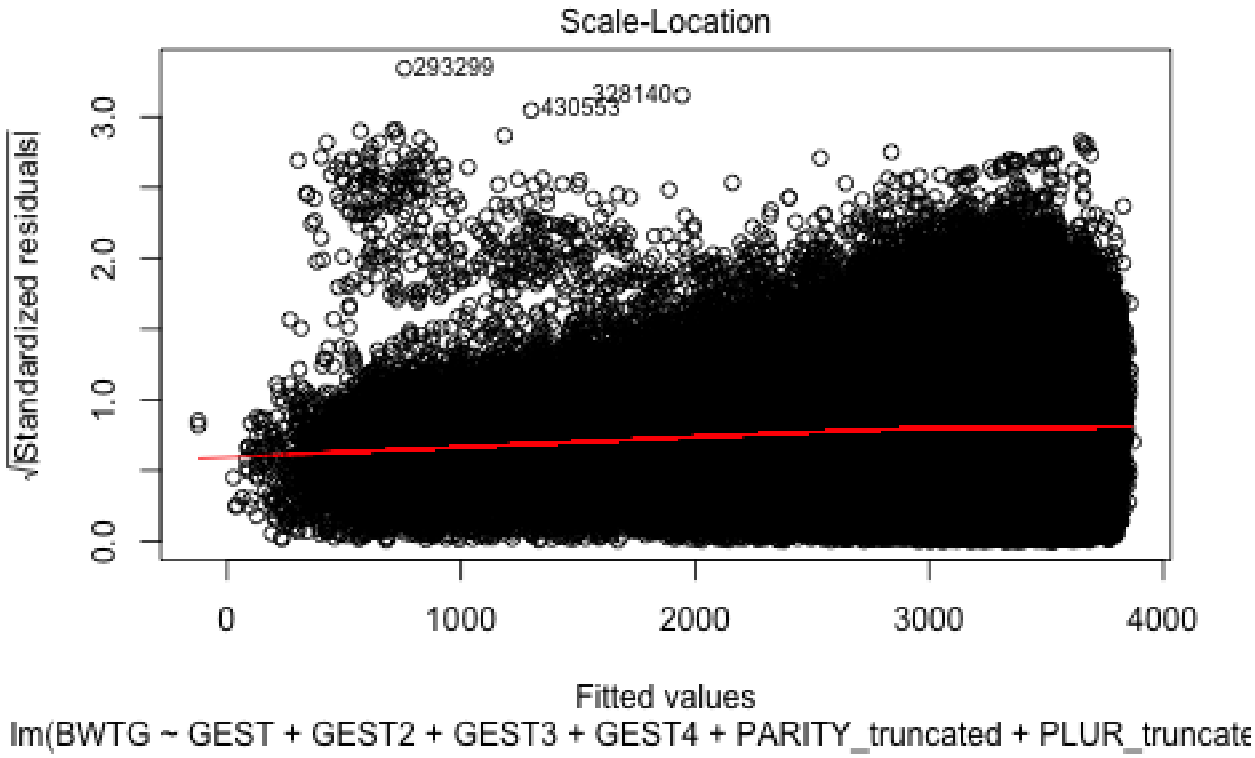
```
model4 = lm(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_truncated
summary(model4)
```

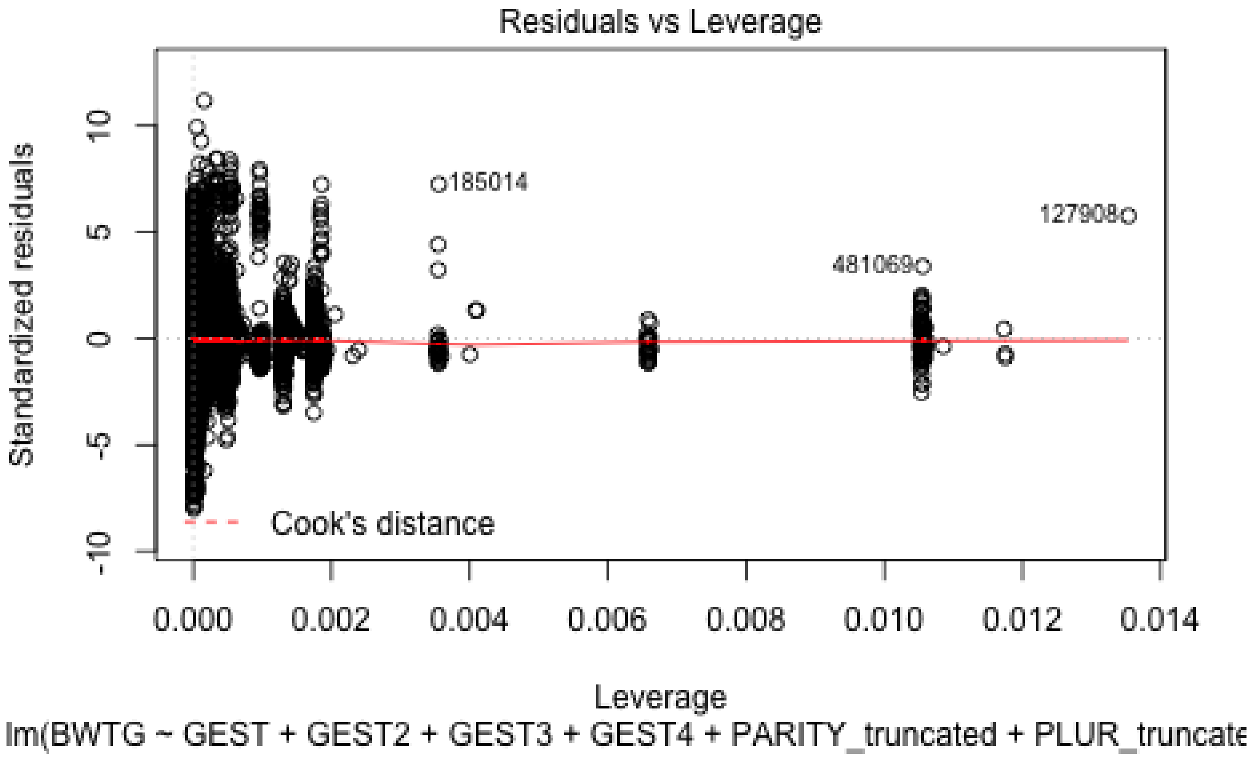
```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3398.1  -280.3   -19.4   257.3  4742.4
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 -1.601e+04  7.646e+02  -20.934  < 2e-16 ***
## GEST                         2.640e+03  1.009e+02   26.157  < 2e-16 ***
## GEST2                       -1.565e+02  4.896e+00  -31.968  < 2e-16 ***
## GEST3                        4.009e+00  1.037e-01   38.660  < 2e-16 ***
## GEST4                       -3.610e-02  8.109e-04  -44.525  < 2e-16 ***
## PARITY_truncated2            8.561e+01  1.308e+00   65.458  < 2e-16 ***
## PARITY_truncated3            1.028e+02  1.528e+00   67.270  < 2e-16 ***
## PARITY_truncated4            1.076e+02  1.912e+00   56.277  < 2e-16 ***
## PARITY_truncated5+           1.104e+02  1.938e+00   56.945  < 2e-16 ***
## PLUR_truncated2              -3.153e+02  2.943e+00 -107.137  < 2e-16 ***
## PLUR_truncated3+             -3.296e+02  1.528e+01  -21.565  < 2e-16 ***
## smoking_typebefore and during -2.024e+02  1.753e+00 -115.467  < 2e-16 ***
## smoking_typebefore only       -1.113e+01  2.676e+00   -4.161 3.17e-05 ***
## smoking_typeduring only       -1.647e+02  9.039e+00  -18.225  < 2e-16 ***
## MAGE                           4.592e+00  9.546e-02   48.105  < 2e-16 ***
## MRACER1                        8.213e+01  1.643e+00   49.981  < 2e-16 ***
## MRACER2                       -9.986e+01  1.803e+00  -55.396  < 2e-16 ***
## MRACER3                       -1.772e+00  4.531e+00   -0.391   0.6958    
## MRACER4                       -4.031e+01  7.527e+00   -5.355 8.56e-08 ***
## MRACER5                       -1.203e+02  1.776e+01   -6.775 1.25e-11 ***
## MRACER6                        2.233e+01  4.359e+01    0.512   0.6086    
## MRACER7                       -2.029e+01  9.365e+00   -2.167   0.0302 *  
## MRACER8                       -1.081e+02  3.177e+00  -34.041  < 2e-16 ***
## mortality                      1.814e+01  3.072e+00    5.905 3.53e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 424.6 on 717845 degrees of freedom
## Multiple R-squared:  0.5279, Adjusted R-squared:  0.5278
## F-statistic: 3.489e+04 on 23 and 717845 DF,  p-value: < 2.2e-16
```

```
plot(model4)
```

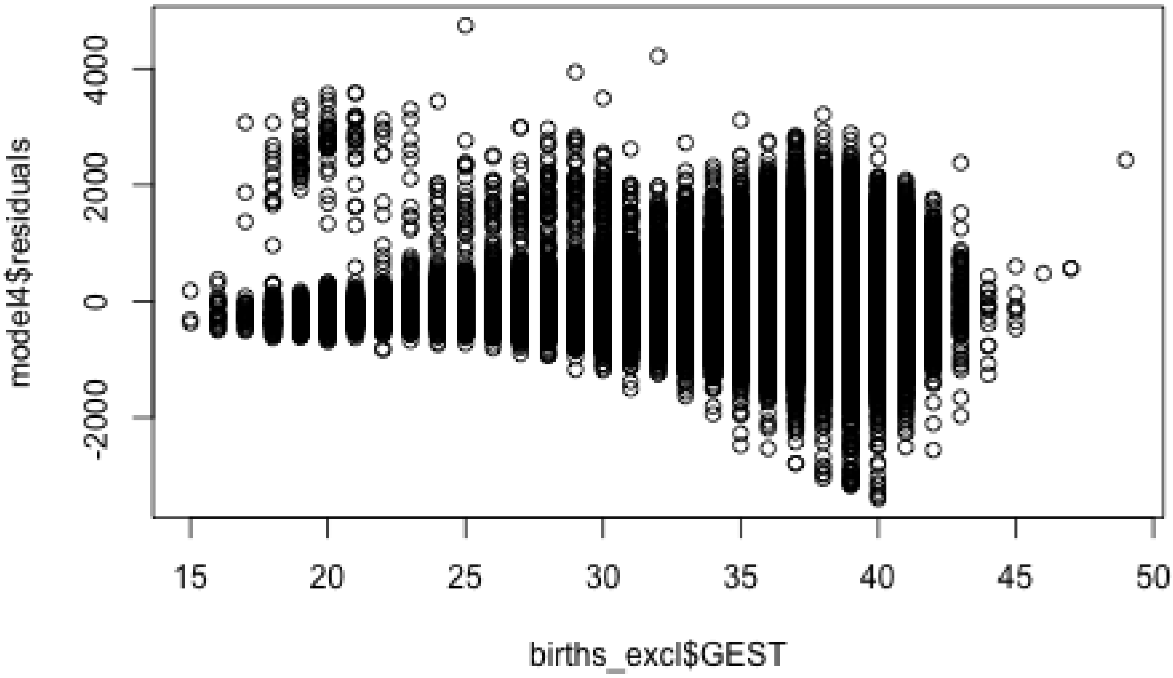



Normal Q-Q
Standardized residuals
10
5
0
-5
-4
-2
0
2
4
Theoretical Quantiles
293299
328140
430553
lm(BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_truncate

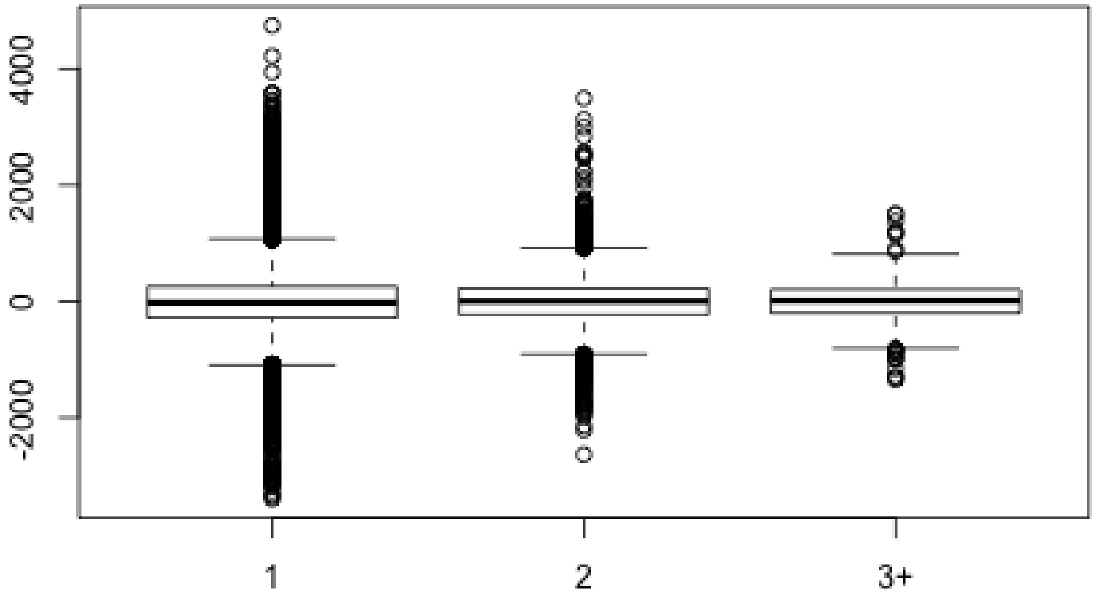

Scale-Location
293299
430553
328140
√|Standardized residuals|
0.0
1.0
2.0
3.0
0
1000
2000
3000
4000
Fitted values
lm(BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_truncate



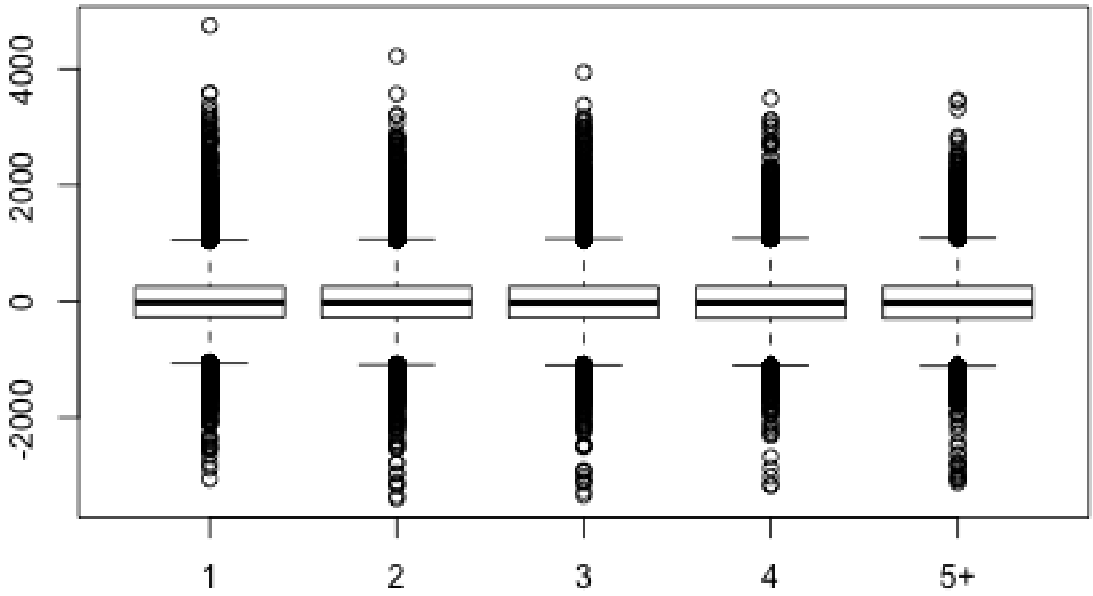
```
# plot(model4$fitted.values, model4$residuals)
plot(births_excl$GEST, model4$residuals)
```



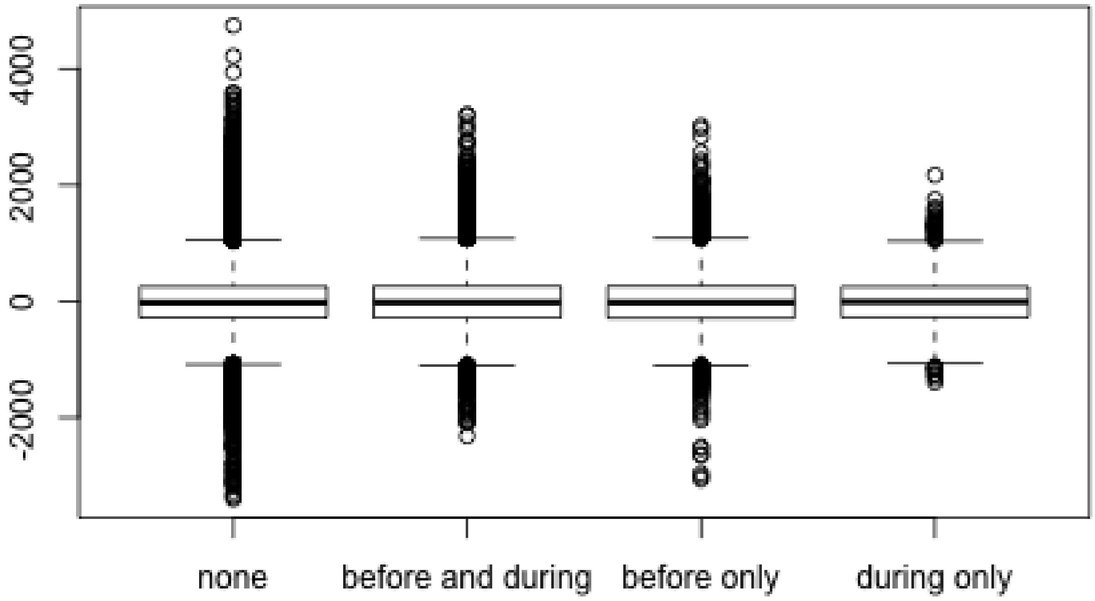
```r
plot(births_excl$PLUR_truncated, model4$residuals)
```



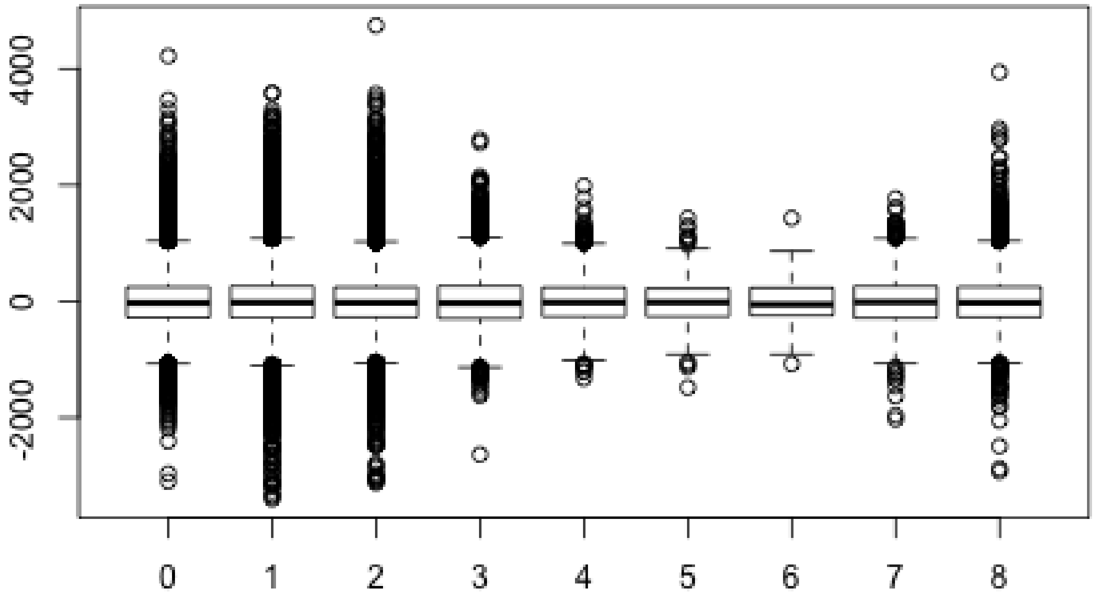
```r
plot(births_excl$PARITY_truncated, model4$residuals)
```



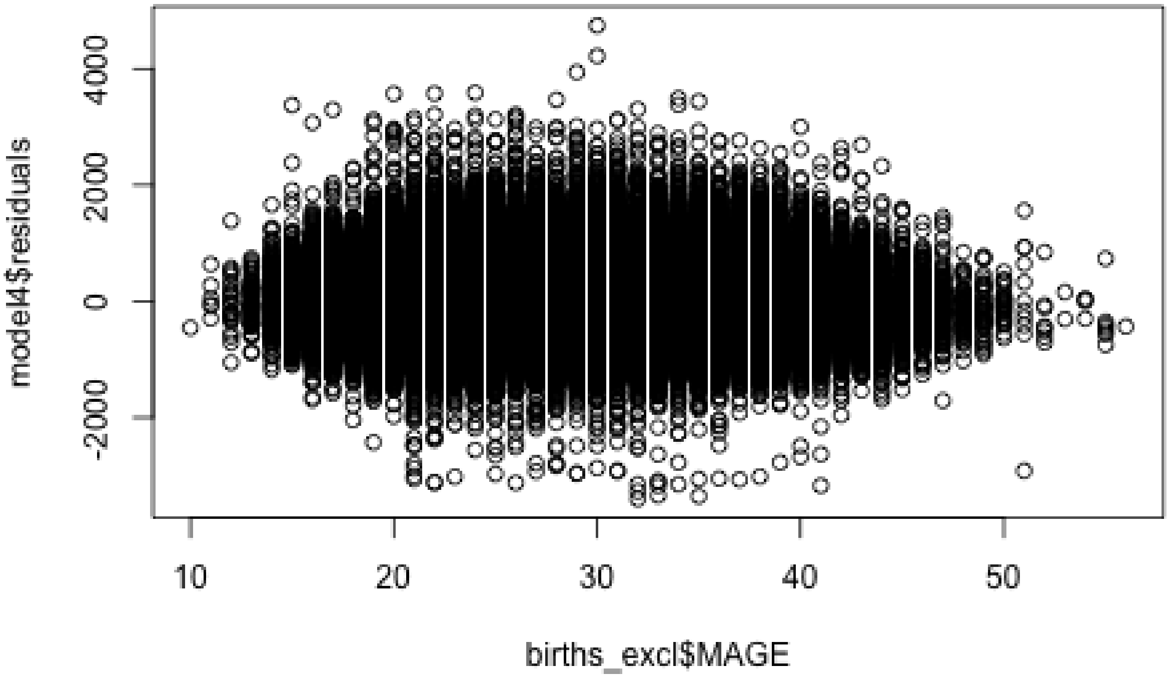
```
plot(births_excl$smoking_type, model4$residuals)
```



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



```r
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.

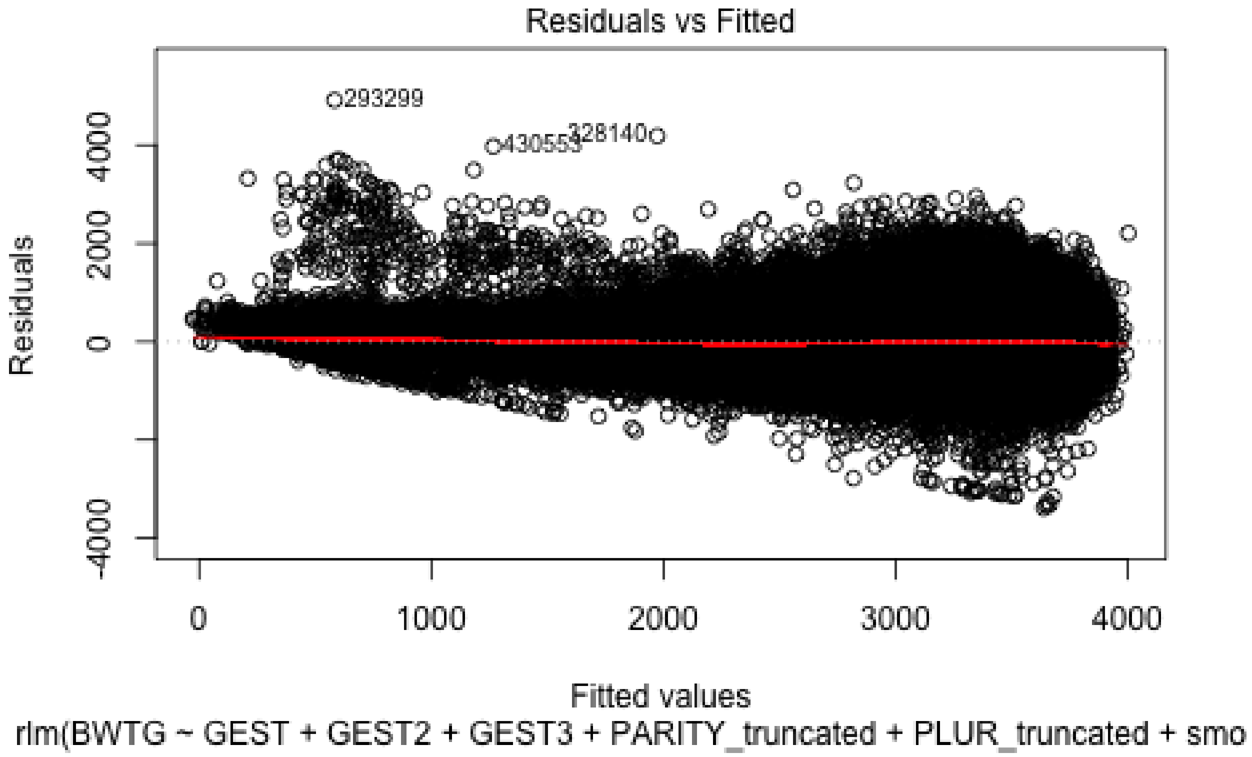
Model 3 looks like the best, but perhaps we can use robust regression to improve upon this the massive residual of the outlier point near 80 weeks of gestational age.

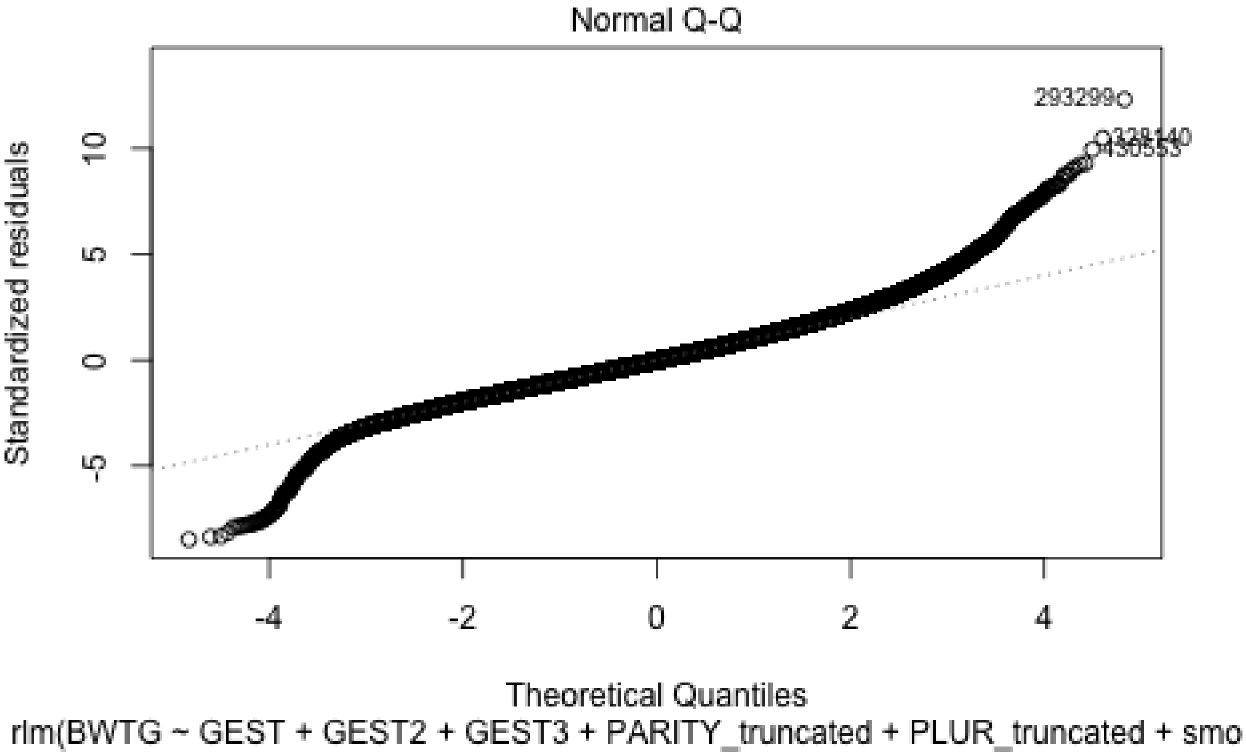
Robust on Model 3

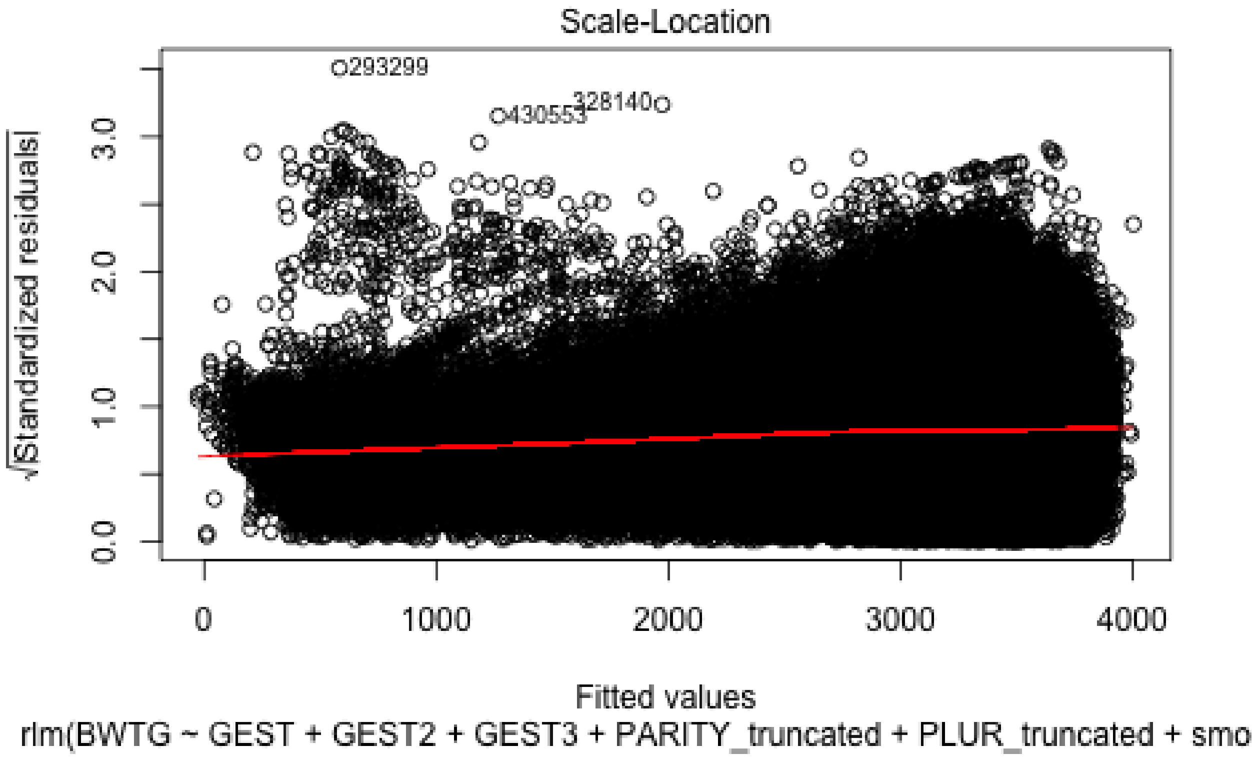
```
robust1 <- rlm(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated + PLUR_truncated + sm
summary(robust1)
```

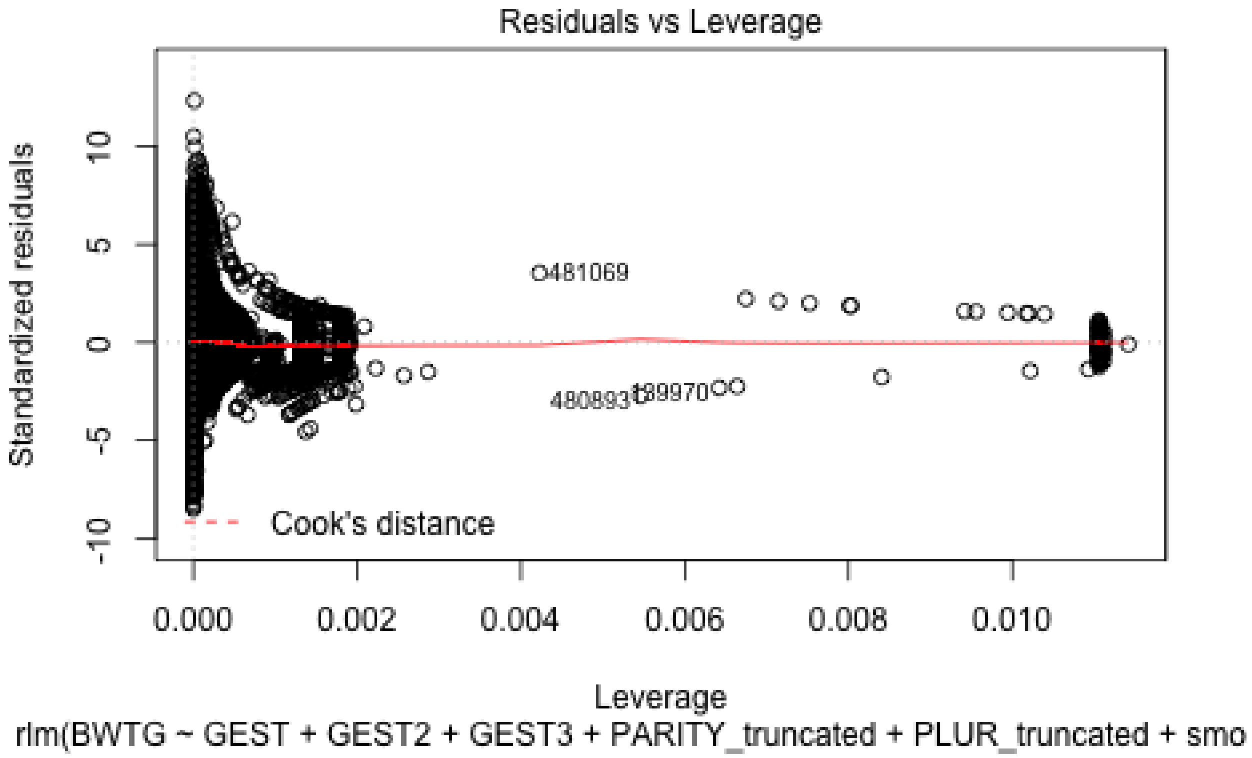
```
##
## Call: rlm(formula = BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
## Residuals:
##       Min        1Q    Median        3Q       Max
## -3387.363  -269.680    -7.516   268.801  4917.081
##
## Coefficients:
##                                 Value      Std. Error t value
## (Intercept)                     15903.5669   182.4842    87.1504
## GEST                            -1690.4719    17.2605   -97.9388
## GEST2                              57.4173     0.5325   107.8355
## GEST3                              -0.5757     0.0054  -107.0707
## PARITY_truncated2                86.8751     1.2796    67.8930
## PARITY_truncated3               104.1254     1.4955    69.6238
## PARITY_truncated4               109.3099     1.8715    58.4092
## PARITY_truncated5+              109.9214     1.8980    57.9151
## PLUR_truncated2                -312.0635     2.8631  -108.9943
## PLUR_truncated3+               -322.9106    14.9615   -21.5828
## smoking_typebefore and during  -202.1493     1.7164  -117.7759
## smoking_typebefore only         -11.6944     2.6205    -4.4627
## smoking_typeduring only        -162.4023     8.8520   -18.3463
## MAGE                              4.4206     0.0935    47.2894
## MRACER1                          85.4898     1.6092    53.1254
## MRACER2                         -98.5591     1.7654   -55.8286
## MRACER3                          -4.5143     4.4368    -1.0175
## MRACER4                         -35.0959     7.3709    -4.7614
## MRACER5                        -112.9427    17.3957    -6.4926
## MRACER6                          27.0106    42.6900     0.6327
## MRACER7                         -15.0170     9.1714    -1.6374
## MRACER8                        -105.6237     3.1111   -33.9504
## mortality                        20.0766     3.0082     6.6740
##
## Residual standard error: 399.2 on 717846 degrees of freedom
```

```
plot(robust1)
```

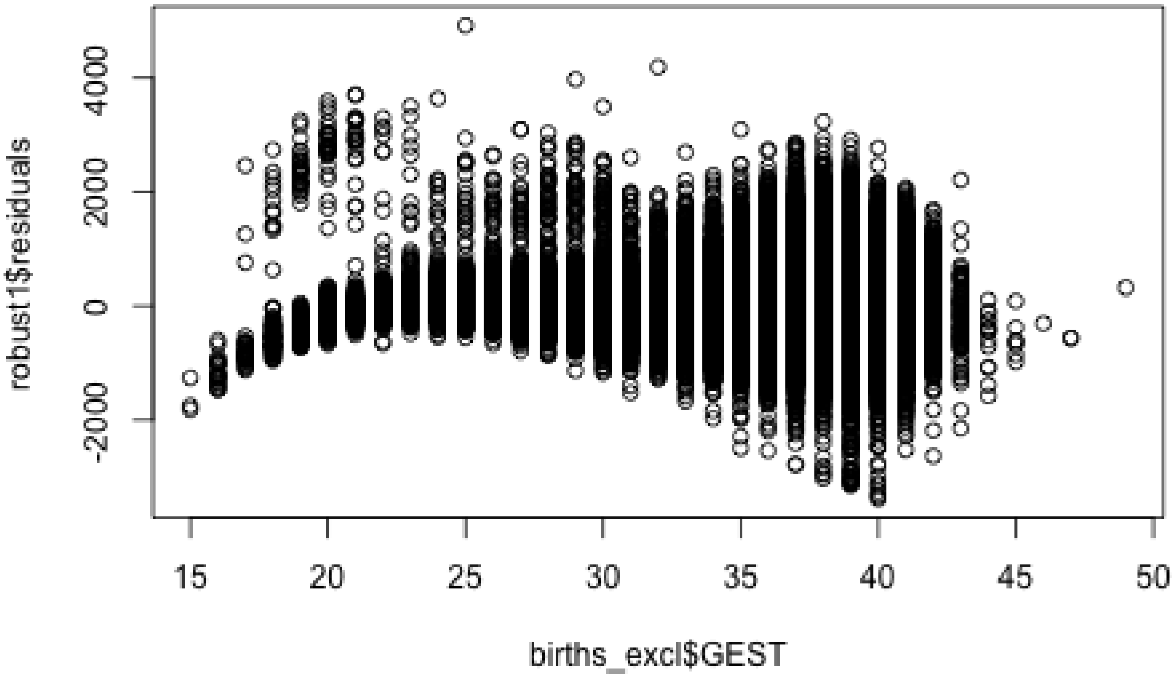



Normal Q-Q
Standardized residuals
10
5
0
-5
-4
-2
0
2
4
293299
328140
430553
Theoretical Quantiles
rlm(BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated + PLUR_truncated + smo

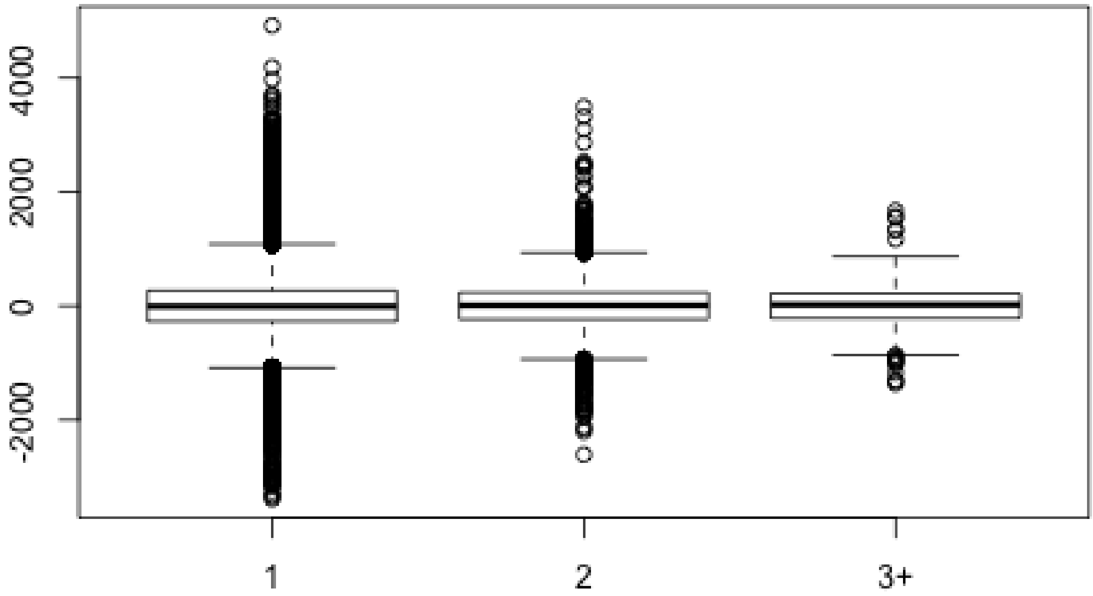

Scale-Location
293299
430553
328140
√|Standardized residuals|
0.0
1.0
2.0
3.0
0
1000
2000
3000
4000
Fitted values
rlm(BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated + PLUR_truncated + smo



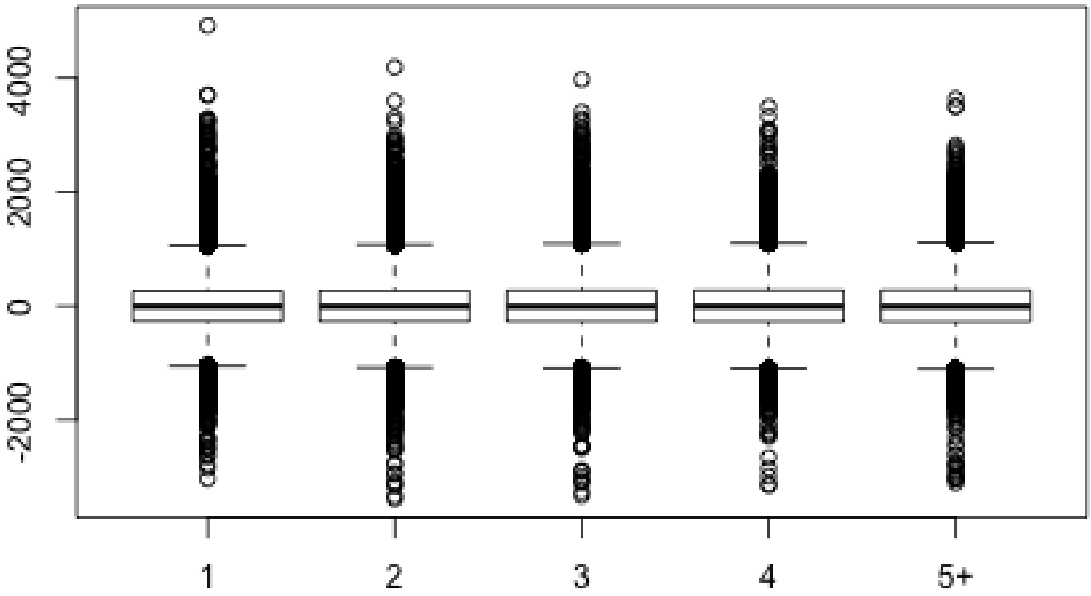
```
# plot(robust1$fitted.values, robust1$residuals)
plot(births_excl$GEST, robust1$residuals)
```



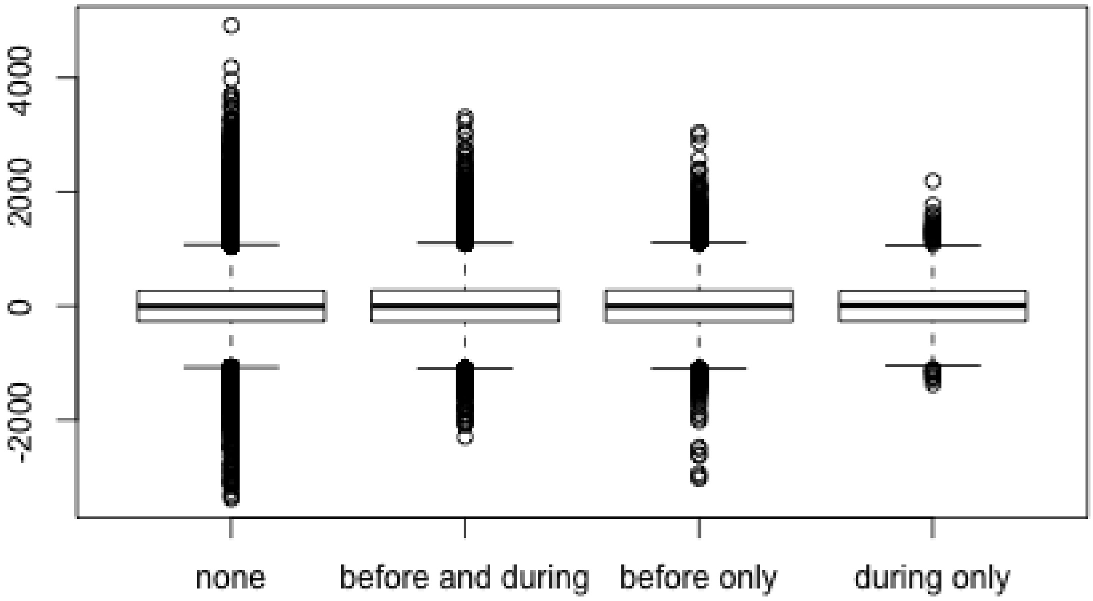
```
plot(births_excl$PLUR_truncated, robust1$residuals)
```



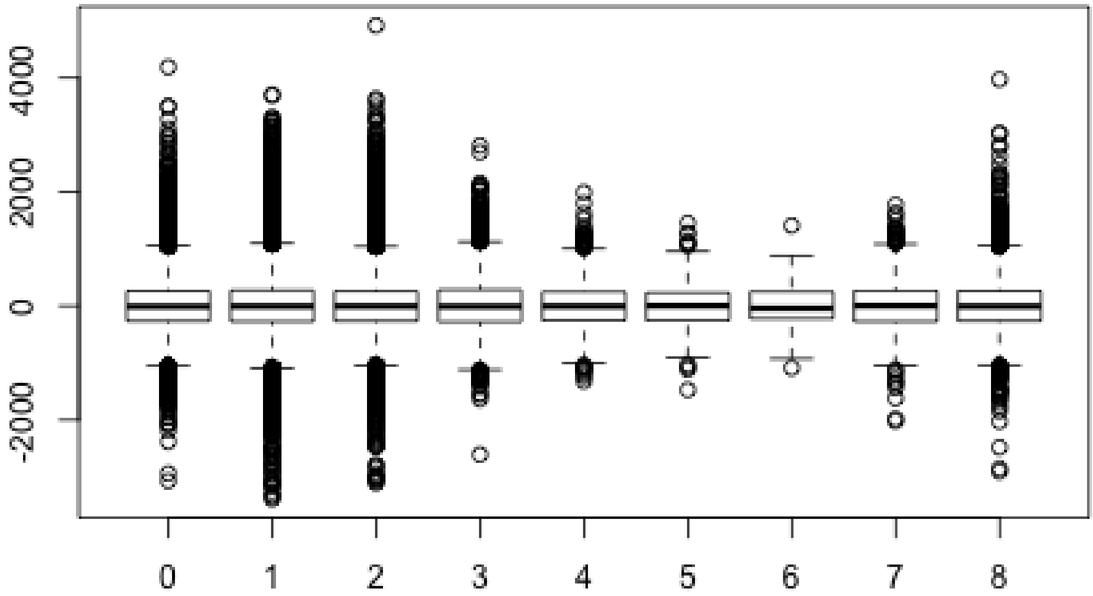
```r
plot(births_excl$PARITY_truncated, robust1$residuals)
```



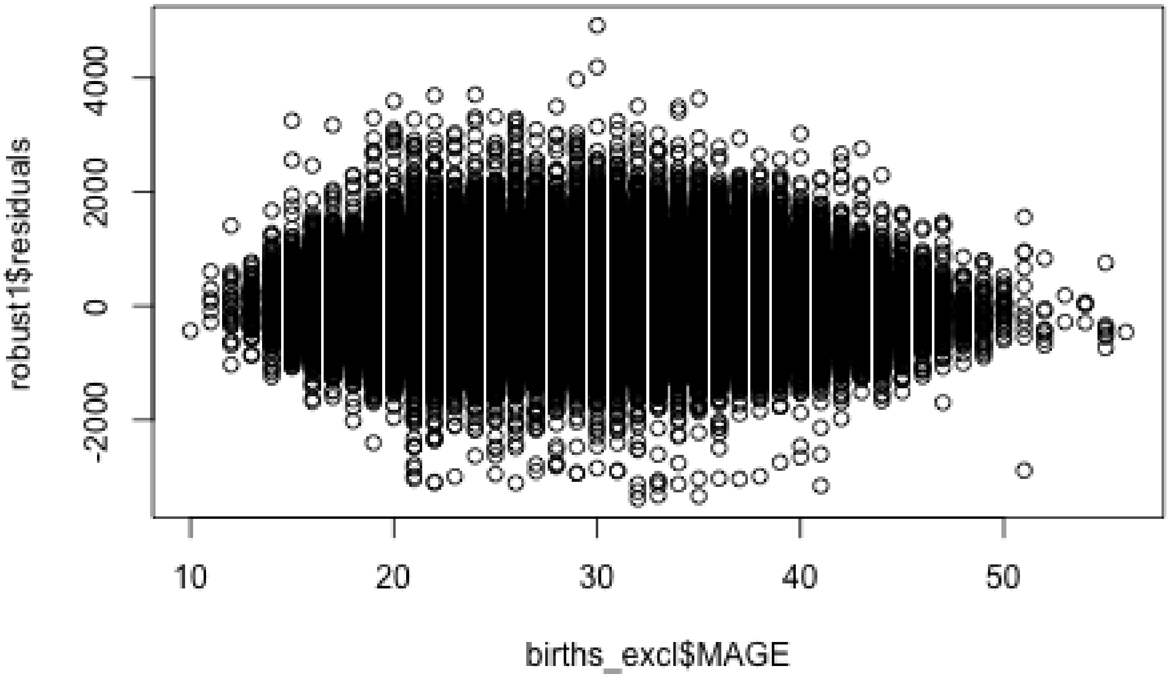
```r
plot(births_excl$smoking_type, robust1$residuals)
```



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



```r
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
## 293299 5500   25 4917.08114 0.1091945
## 328140 6160   32 4186.78682 0.1282402
## 430553 5239   29 3971.68803 0.1351850
## 98978  4309   21 3704.98342 0.1449236
## 48677  4281   21 3691.28387 0.1454614
## 717861 3289   39   45.42608 1.0000000
## 717862 3770   41  106.21740 1.0000000
## 717863 2872   39 -371.57392 1.0000000
## 717864 3360   38   54.06698 1.0000000
## 717866 3713   40  194.44419 1.0000000
## 717867 3180   39 -283.40570 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).