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)
library(quantreg)
knitr::opts_chunk$set(warning=FALSE)
```

Load and Clean Data

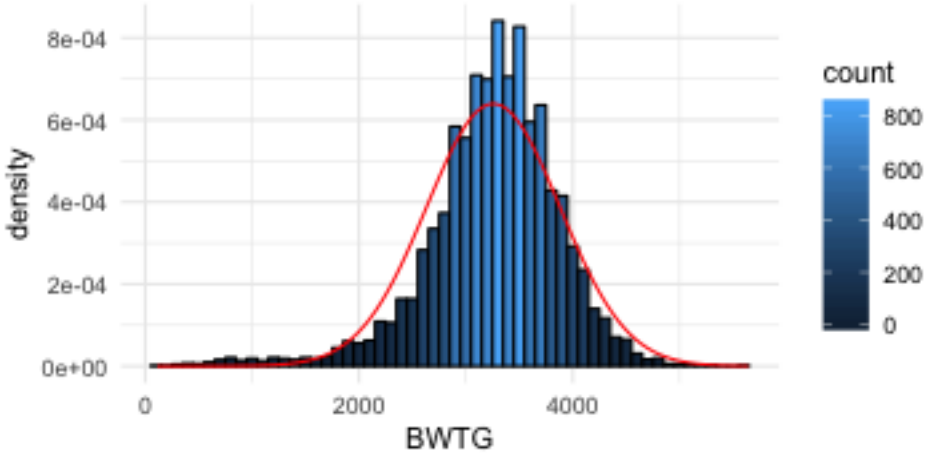
```
## read in data
births = read.csv("data/Yr1116Birth.csv", na.strings = "9999")
deaths = read.csv("data/Yr1116Death.csv")

## 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
```

Make smaller subset – take out in final version, just for formatting

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

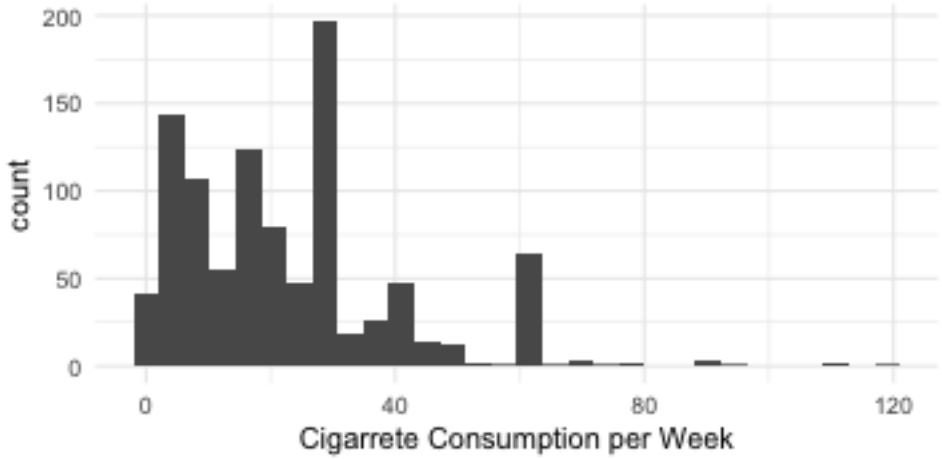
Exploratory Data Analysis

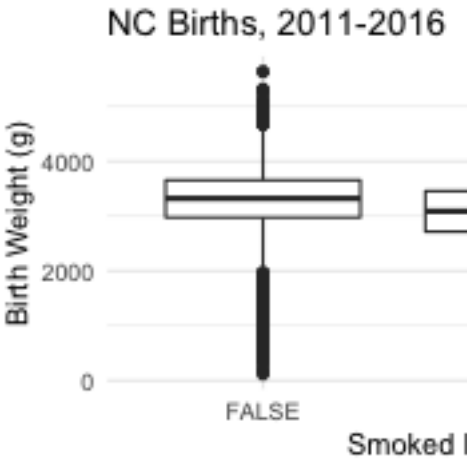
Birthweight

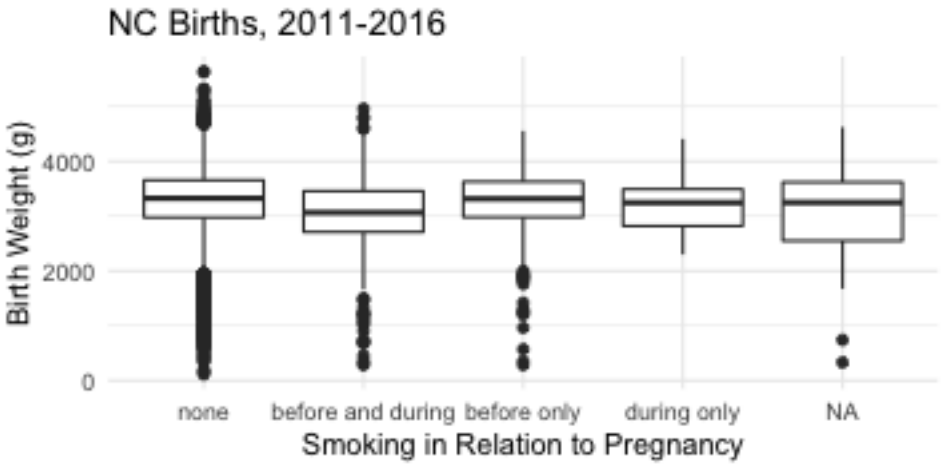


Birthweight is close to normally distrubted, with a slight left skew, centered around ~3300g with a standard deviation of 600g. There appear to be no large outliers in terms of birthweight. 430 birth weights are missing. We see that the left tail is much larger than we would expect in a normal distribution

Smoking

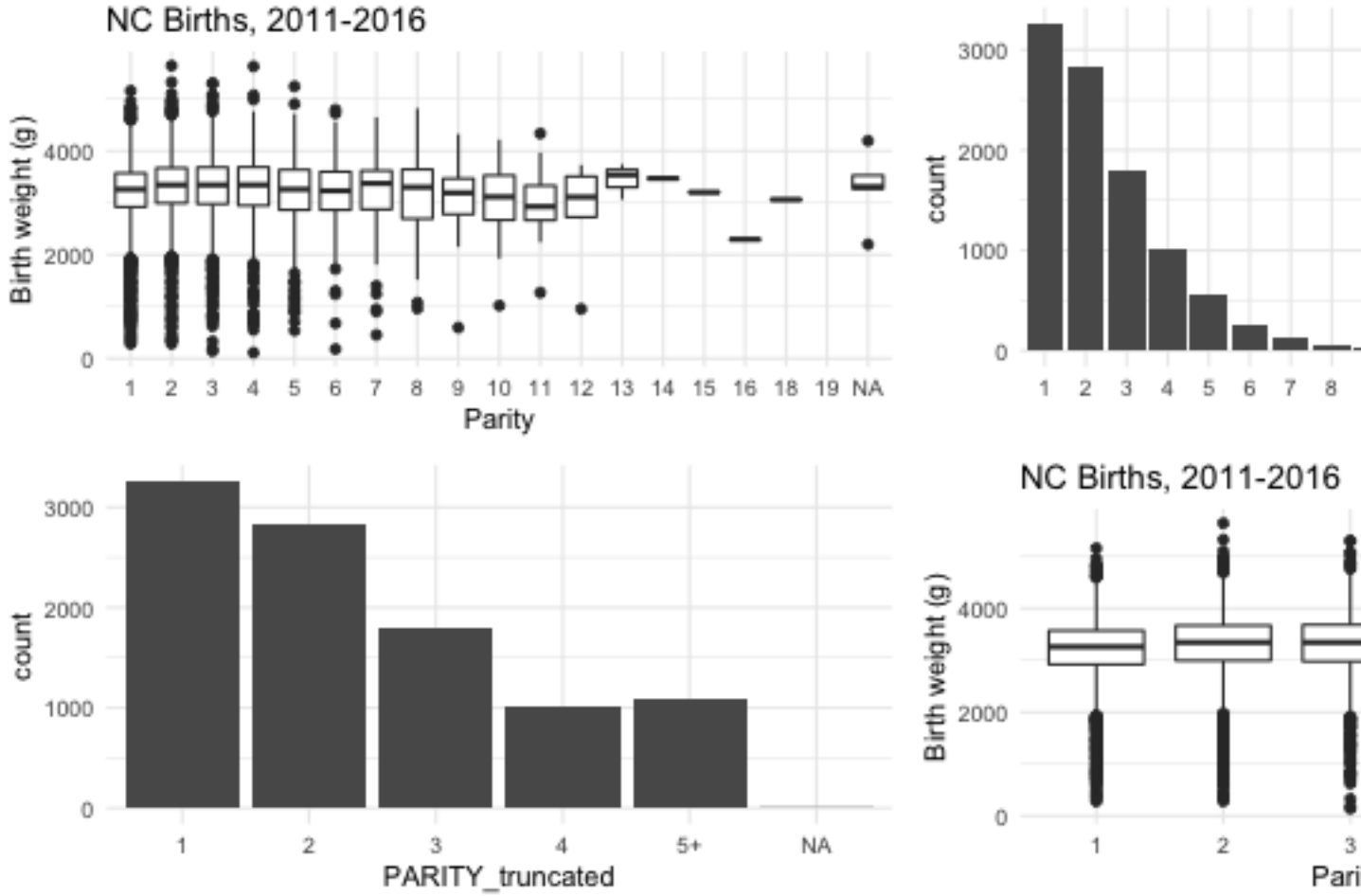






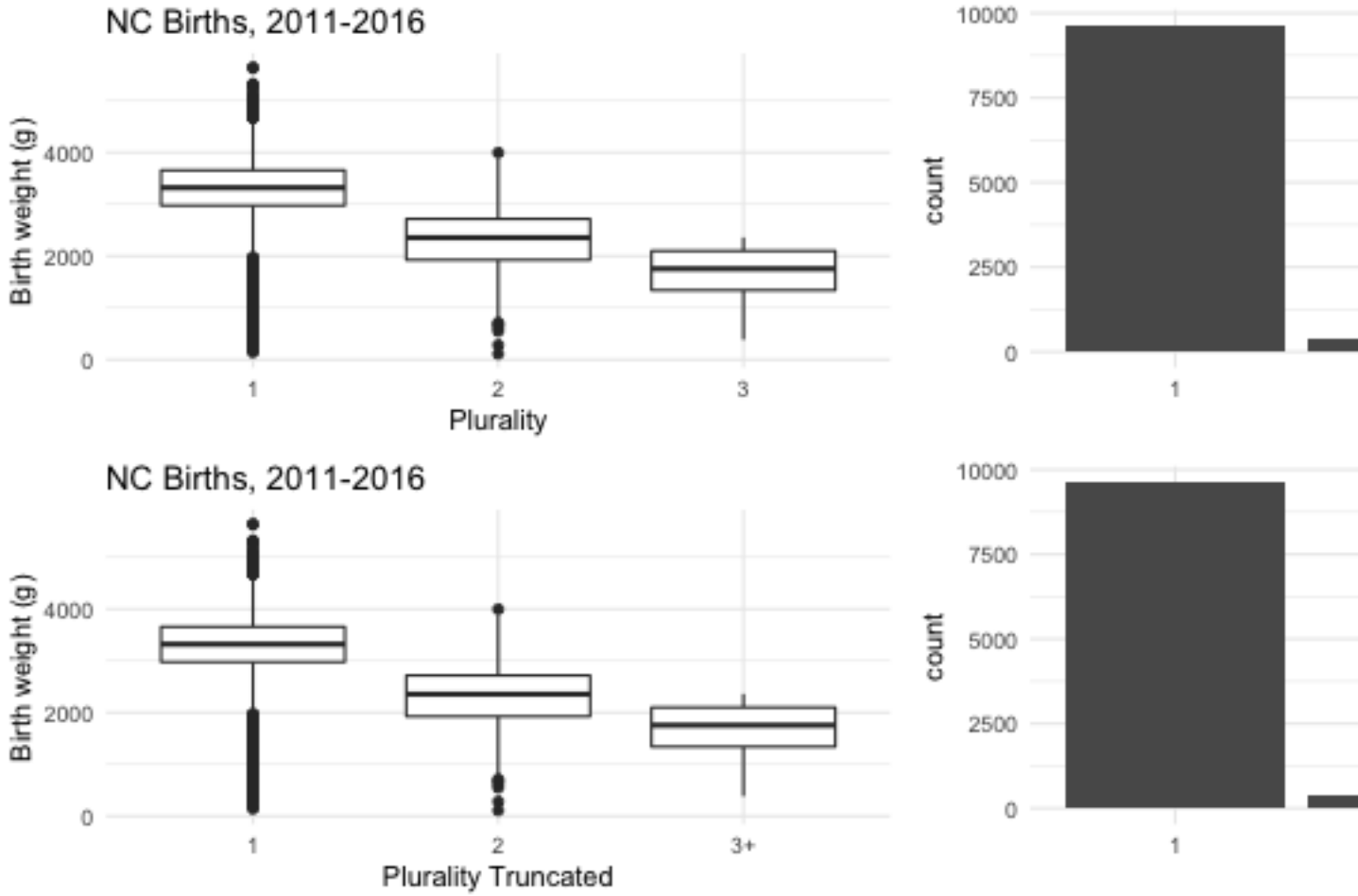
Around 13% of women smoked in the three months leading up to pregnancy and around 10% of women at any point during their pregnancy. Among those who did smoke during pregnance, the average number of cigarettes smoked during pregnancy was 23. The birthweight of children of smokers was significantly lower 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



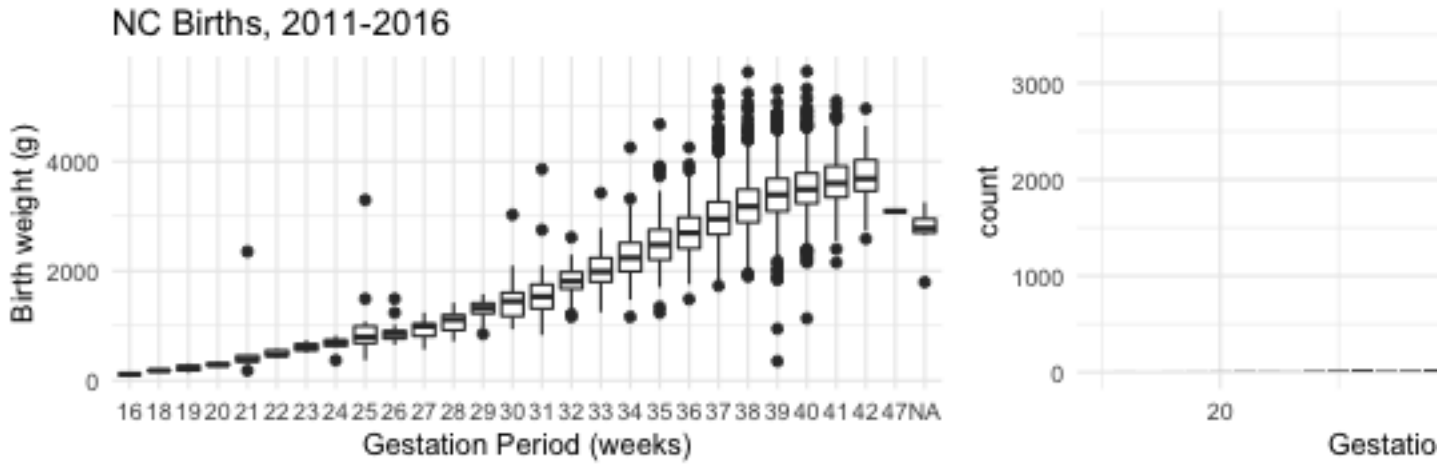
Independent of other variables, we see a negative relationship between parity and birth weight past the first child. The frequency of parity decreases in an exponential fashion. A second variable was created that truncates parities of at least five to improve interprability and prevent overfitting. The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

Plurality



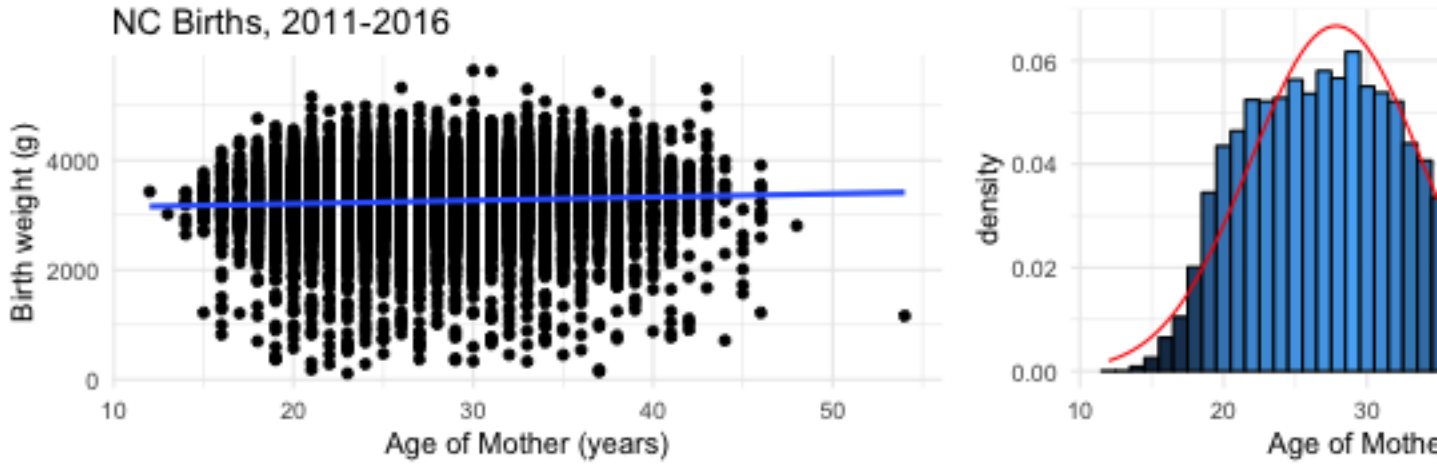
We see a strong non linear negative relationship between plurality and birth weight. The frequency of pluralities above two is extremely small, and we again see a proportionally small amount of missing data. A second variable was created that truncates pluralities of at least three to improve interprability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

Gestation



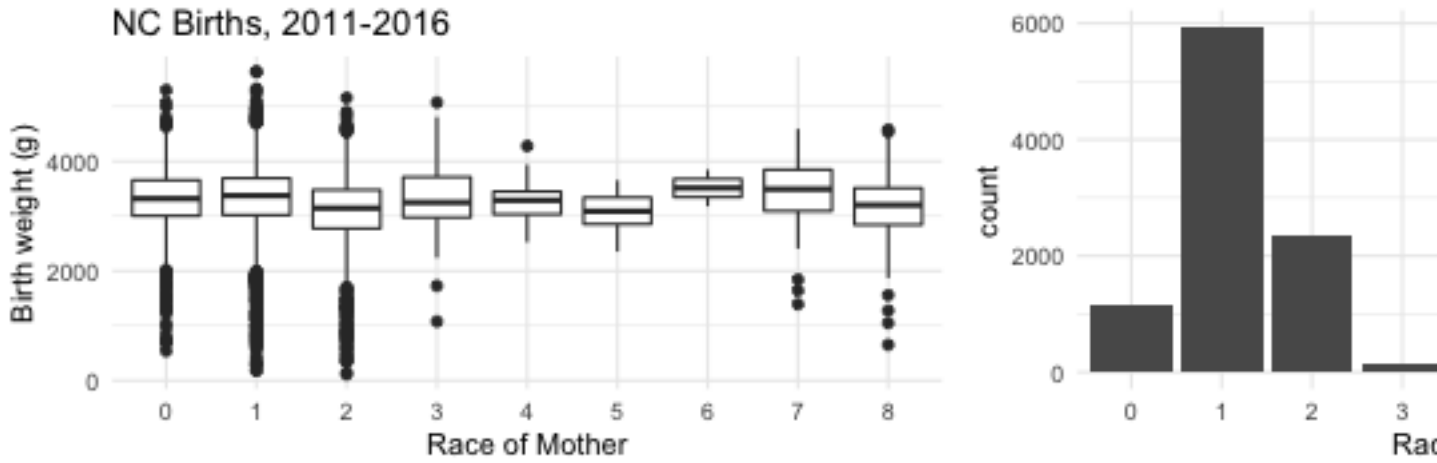
There appears to be a non linear positive relationship between gestation period and birth weight. The mean gestational period is approximately 38.5 weeks and the period with the highest median weight is 42 weeks. The frequency distribution is left skewed with the majority of babies having a gestational 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



Mother's age seems to be fairly normally distributed with a mean of 27.7. 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.

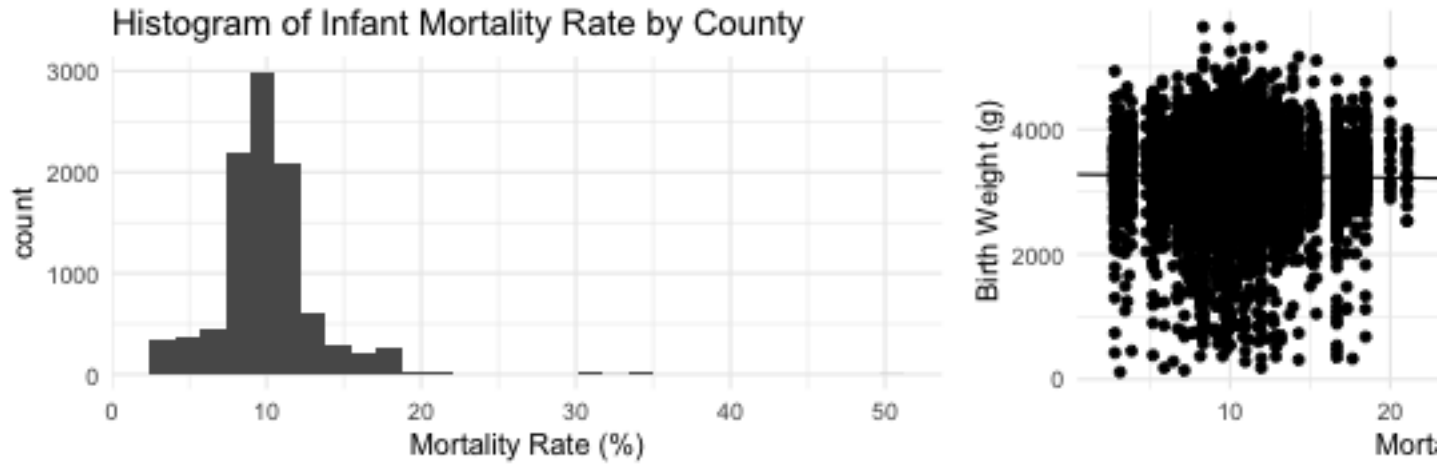
Race of Mother



0 - Other non-White 1 - White 2 - Black or African American 3 - American Indian or Alaska Native 4 - Chinese 5 - Japanese 6 - Native Hawaiian 7 - Filipino 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 other non-white, and 3 percent identify as other asian.

County / Socioeconomic Status



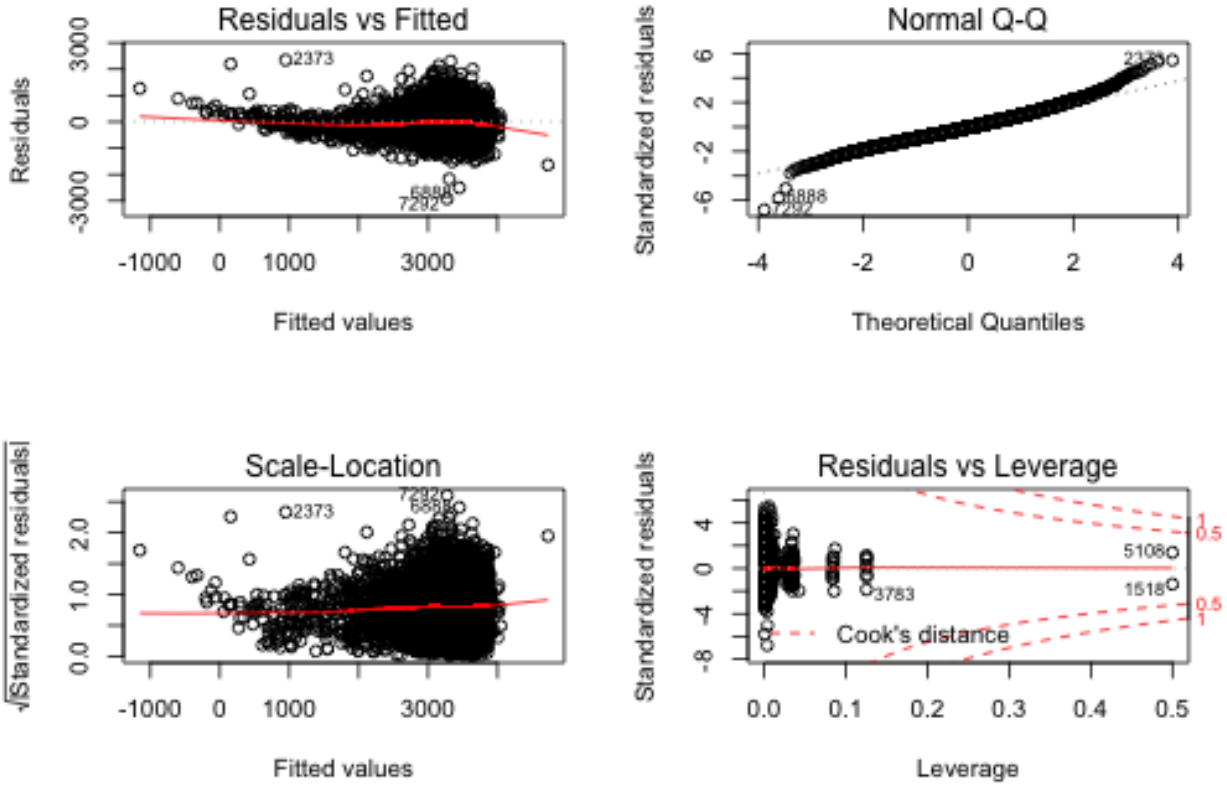
We chose to use infant mortality rate of birth county as a proxy for socioeconomic status, calculated as number of deaths before the age of 1 divided by total number of births in a county. The median county in the data had a infant mortality rate of 0.7%, with the range of infant mortality rates in our 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

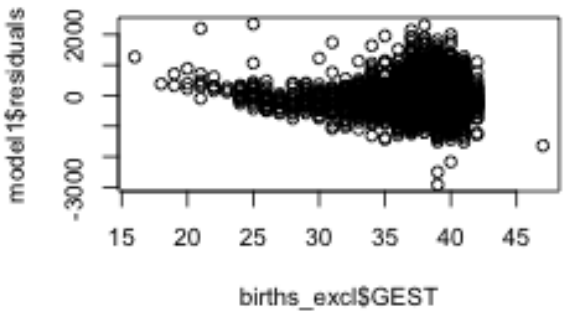
```
births_excl = na.omit(births)
births_excl = births_excl[which(births_excl$GEST < 80), ]
births_excl = births_excl %>%
  mutate(GEST2 = GEST^2, GEST3 = GEST ^ 3, GEST4 = GEST^4)
model1 = lm(data = births_excl, BWTG ~ GEST + PARITY_truncated + PLUR_truncated + smoking_type + MAGE +
summary(model1)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

```
##
## 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
## -2922.23  -286.95   -20.77   263.85  2345.22
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3736.2320    83.3533 -44.824  < 2e-16 ***
## GEST                           176.8419     2.0131  87.845  < 2e-16 ***
## PARITY_truncated2              102.8395    11.4110   9.012  < 2e-16 ***
## PARITY_truncated3              131.0085    13.2889   9.858  < 2e-16 ***
## PARITY_truncated4              131.5831    16.3297   8.058 8.67e-16 ***
## PARITY_truncated5+             114.0909    16.5640   6.888 6.01e-12 ***
## PLUR_truncated2               -371.8108    25.0819 -14.824  < 2e-16 ***
## PLUR_truncated3+              -497.1656   125.5292  -3.961 7.53e-05 ***
## smoking_typebefore and during -197.4210    15.3091 -12.896  < 2e-16 ***
## smoking_typebefore only        -29.3014    23.0530  -1.271   0.2037
## smoking_typeduring only       -130.0501    80.5905  -1.614   0.1066
## MAGE                             3.9843     0.8152   4.888 1.04e-06 ***
## MRACER1                         70.2371    14.2035   4.945 7.74e-07 ***
## MRACER2                       -101.9657    15.7055  -6.492 8.86e-11 ***
## MRACER3                         52.7899    40.6045   1.300   0.1936
## MRACER4                       -105.8141    74.2535  -1.425   0.1542
## MRACER5                       -284.3352   153.5007  -1.852   0.0640 .
## MRACER6                        220.7664   305.9593   0.722   0.4706
## MRACER7                        159.8826    80.0729   1.997   0.0459 *
## MRACER8                       -128.5909    27.8139  -4.623 3.83e-06 ***
## mortality                        1.5905     1.3053   1.219   0.2231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432.2 on 9818 degrees of freedom
## Multiple R-squared:  0.5201, Adjusted R-squared:  0.5192
## F-statistic: 532.1 on 20 and 9818 DF,  p-value: < 2.2e-16
```

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




```
plot(births_excl$GEST, model1$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.

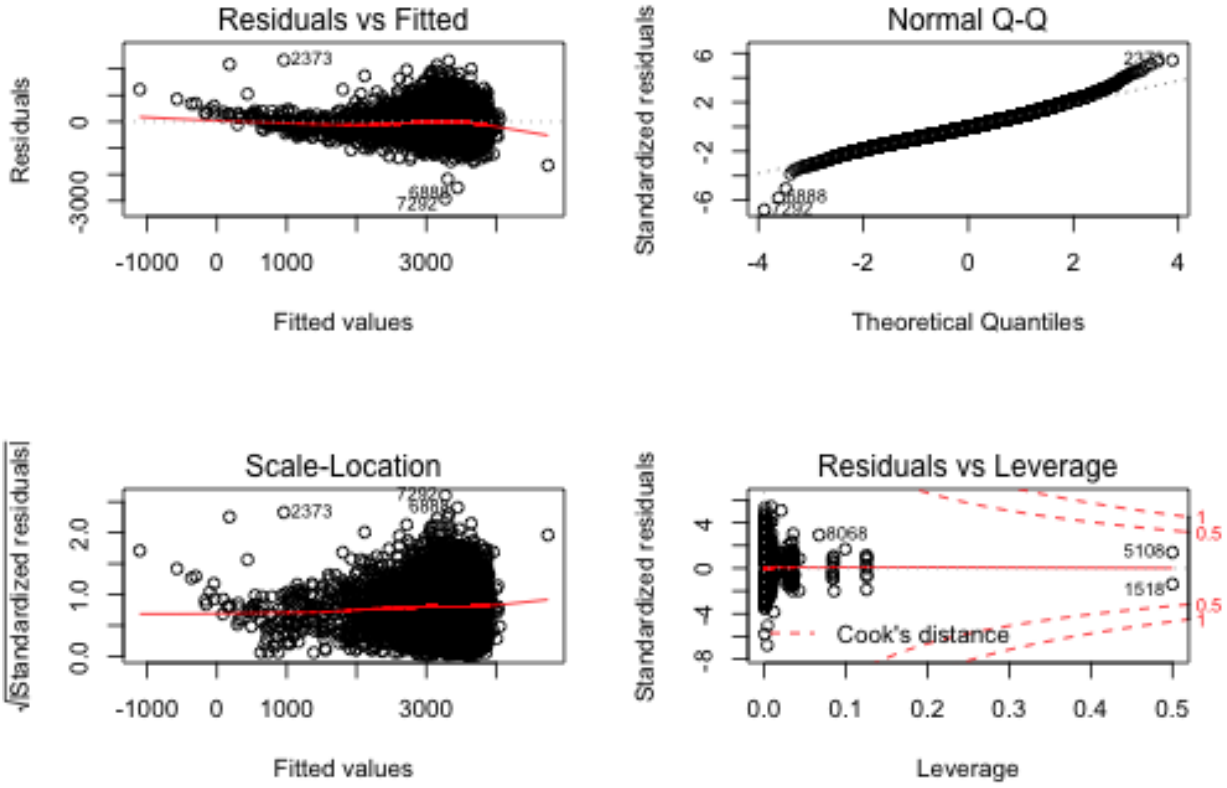
The residual graph for Mother's race indicates that residuals are lower for for races 3, 4, 5, 6, and 7 and higher for the other races. This could be something to explore.

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).

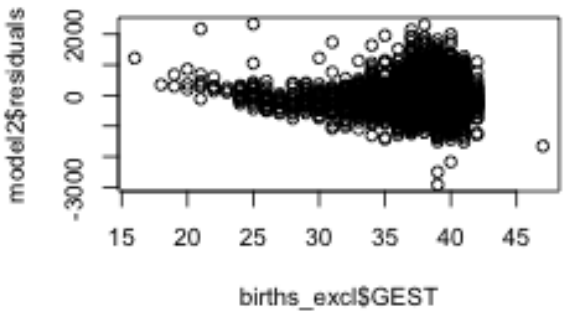
```
model2 = lm(data = births_excl, BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated + smoking_type +
summary(model2)
```

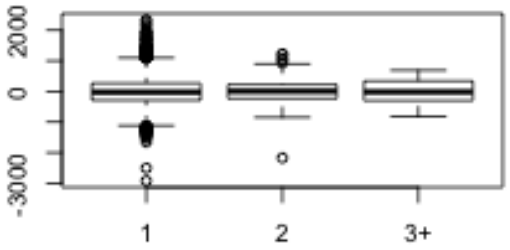
```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated +
##     smoking_type + MAGE + MRACER + mortality, data = births_excl,
##     na.action = "na.exclude")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2922.2  -287.6   -20.8   262.4  2335.1
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3578.1822   362.5317  -9.870  < 2e-16 ***
## GEST                           167.5662    20.8040   8.055 8.91e-16 ***
## GEST2                            0.1335     0.2980   0.448   0.6542
## PARITY_truncated2              103.0916    11.4254   9.023  < 2e-16 ***
## PARITY_truncated3              131.3741    13.3145   9.867  < 2e-16 ***
## PARITY_truncated4              131.9032    16.3460   8.069 7.89e-16 ***
## PARITY_truncated5+             114.5321    16.5940   6.902 5.44e-12 ***
## PLUR_truncated2               -370.7776    25.1887 -14.720  < 2e-16 ***
## PLUR_truncated3+              -497.8337   125.5432  -3.965 7.38e-05 ***
## smoking_typebefore and during -197.3202    15.3114 -12.887  < 2e-16 ***
## smoking_typebefore only        -29.2771    23.0540  -1.270   0.2041
## smoking_typeduring only       -129.8809    80.5947  -1.612   0.1071
## MAGE                             3.9871     0.8152   4.891 1.02e-06 ***
## MRACER1                         70.2770    14.2044   4.948 7.64e-07 ***
## MRACER2                       -101.8882    15.7071  -6.487 9.19e-11 ***
## MRACER3                         52.8616    40.6065   1.302   0.1930
## MRACER4                       -105.9045    74.2568  -1.426   0.1538
## MRACER5                       -284.0352   153.5084  -1.850   0.0643 .
## MRACER6                        221.5041   305.9762   0.724   0.4691
## MRACER7                        160.0520    80.0770   1.999   0.0457 *
## MRACER8                       -128.3413    27.8206  -4.613 4.02e-06 ***
## mortality                        1.5990     1.3055   1.225   0.2207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432.2 on 9817 degrees of freedom
## Multiple R-squared:  0.5202, Adjusted R-squared:  0.5191
## F-statistic: 506.8 on 21 and 9817 DF,  p-value: < 2.2e-16
```

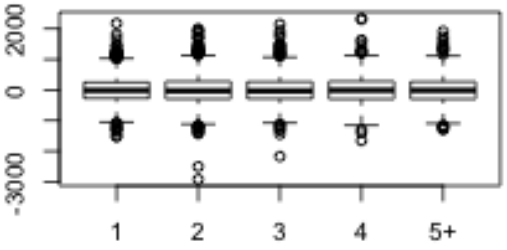
```
par(mfrow = c(2,2))
plot(model2)
```

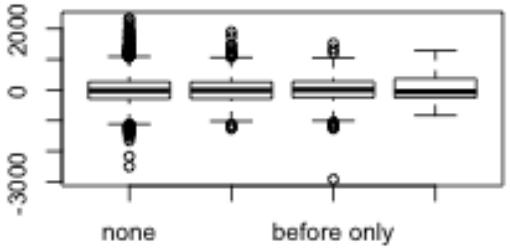



```
# plot(model2$fitted.values, model2$residuals)
plot(births_excl$GEST, model2$residuals)
plot(births_excl$PLUR_truncated, model2$residuals)
plot(births_excl$PARITY_truncated, model2$residuals)
plot(births_excl$smoking_type, model2$residuals)
```

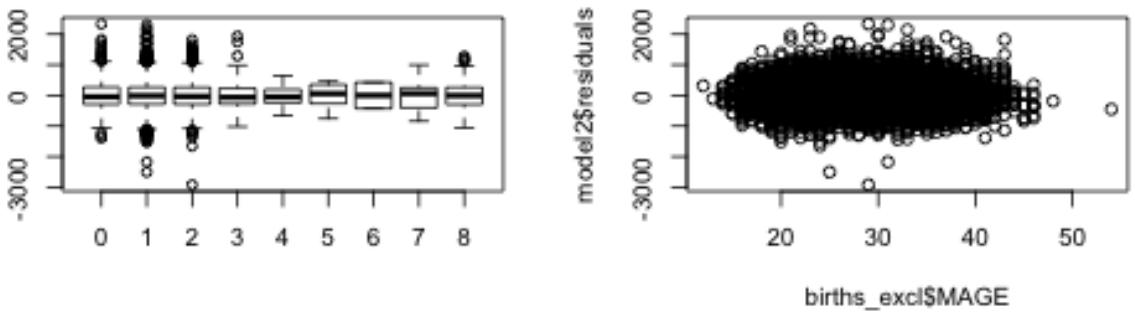









```
plot(births_excl$MRACER, model2$residuals)
plot(births_excl$MAGE, model2$residuals)
```



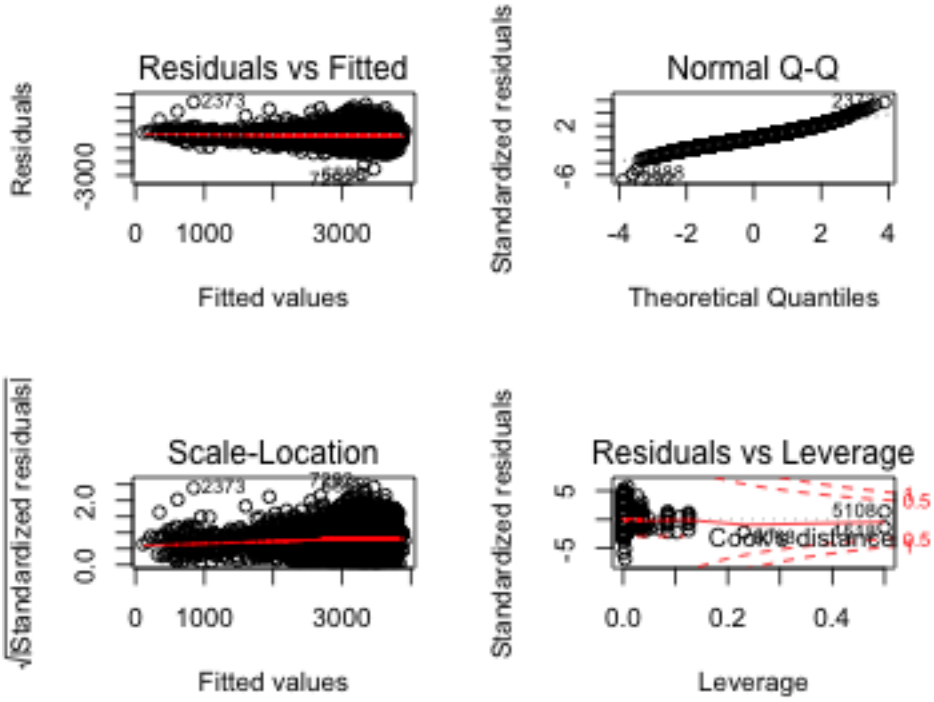
The new model still displays the original downwards trend in the residual vs gestational period graph. Perhaps another transformation on GEST would be helpful – a cubic term can be added. The other residuals plots also retain their trends from model 1.

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

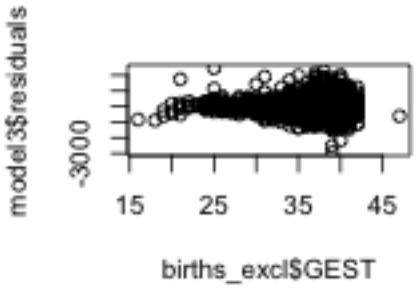
```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2935.58  -285.66   -20.19   259.71  2449.80
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.526e+04  1.631e+03   9.354  < 2e-16 ***
## GEST                        -1.623e+03  1.526e+02 -10.633  < 2e-16 ***
## GEST2                        5.533e+01  4.671e+00  11.845  < 2e-16 ***
## GEST3                       -5.551e-01  4.688e-02 -11.840  < 2e-16 ***
## PARITY_truncated2            9.488e+01  1.137e+01   8.348  < 2e-16 ***
## PARITY_truncated3            1.218e+02  1.325e+01   9.198  < 2e-16 ***
## PARITY_truncated4            1.228e+02  1.625e+01   7.557 4.48e-14 ***
```

```
## PARITY_truncated5+             1.082e+02  1.649e+01   6.563 5.54e-11 ***
## PLUR_truncated2               -3.440e+02  2.511e+01 -13.695  < 2e-16 ***
## PLUR_truncated3+              -4.409e+02  1.248e+02  -3.534 0.000411 ***
## smoking_typebefore and during -1.987e+02  1.520e+01 -13.065  < 2e-16 ***
## smoking_typebefore only       -3.218e+01  2.289e+01  -1.406 0.159822
## smoking_typeduring only       -1.341e+02  8.003e+01  -1.675 0.093910 .
## MAGE                           4.037e+00  8.095e-01   4.987 6.25e-07 ***
## MRACER1                        6.753e+01  1.411e+01   4.787 1.72e-06 ***
## MRACER2                       -1.046e+02  1.560e+01  -6.704 2.13e-11 ***
## MRACER3                        4.915e+01  4.032e+01   1.219 0.222923
## MRACER4                       -1.094e+02  7.374e+01  -1.484 0.137772
## MRACER5                       -2.957e+02  1.524e+02  -1.940 0.052389 .
## MRACER6                        1.991e+02  3.038e+02   0.655 0.512294
## MRACER7                        1.718e+02  7.952e+01   2.161 0.030755 *
## MRACER8                       -1.356e+02  2.763e+01  -4.909 9.32e-07 ***
## mortality                      1.746e+00  1.296e+00   1.347 0.178069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 429.2 on 9816 degrees of freedom
## Multiple R-squared:  0.5269, Adjusted R-squared:  0.5259
## F-statistic: 496.9 on 22 and 9816 DF,  p-value: < 2.2e-16
```

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




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



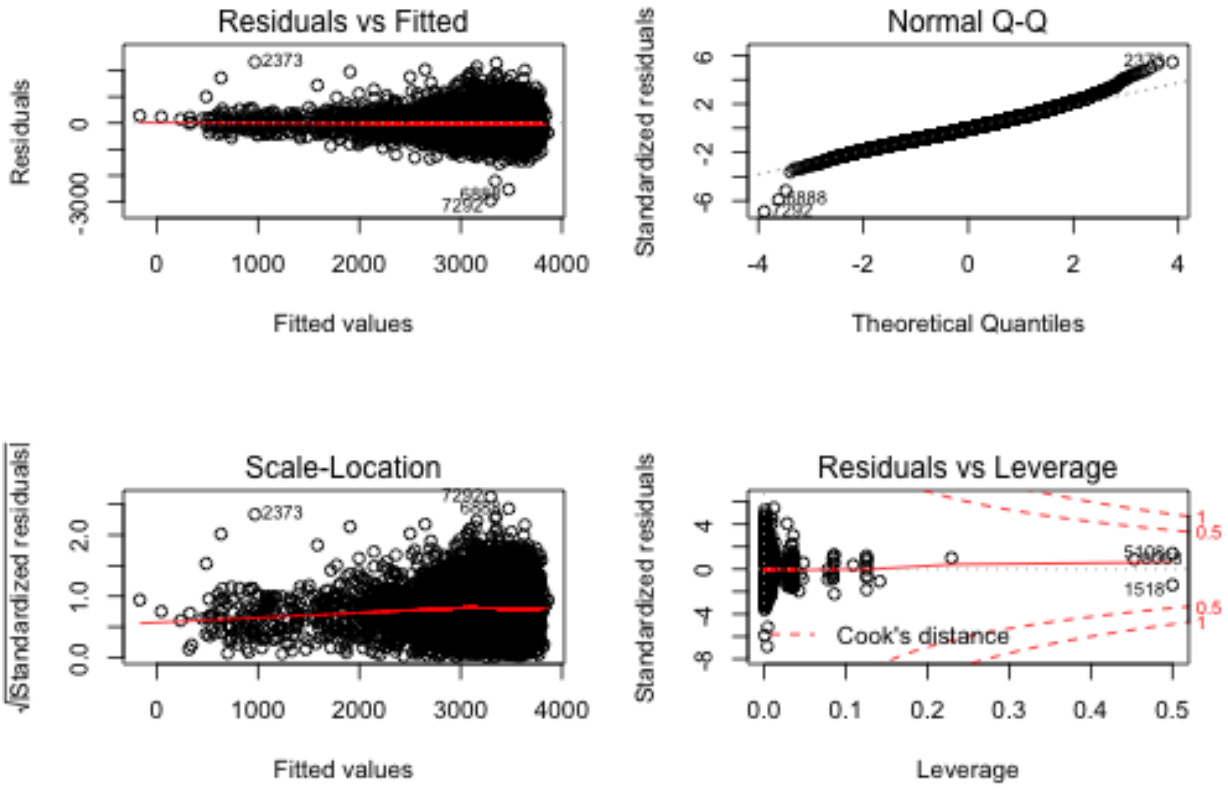
The residual vs gestational period shows a much more random pattern than before. It is worth investigating if adding a quartic term would help. The other residuals plots also retain their trends from model 1.

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

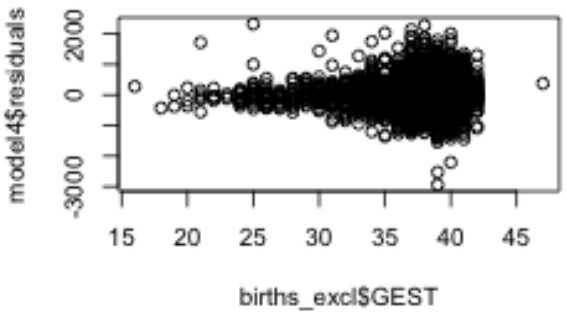
```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2943.30  -285.52   -21.33   258.64  2325.06
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.626e+04  5.930e+03  -2.742 0.006127 **
## GEST                            2.588e+03  7.771e+02   3.331 0.000869 ***
## GEST2                          -1.503e+02  3.750e+01  -4.009 6.15e-05 ***
## GEST3                           3.811e+00  7.913e-01   4.816 1.49e-06 ***
## GEST4                          -3.410e-02  6.169e-03  -5.527 3.34e-08 ***
## PARITY_truncated2               9.304e+01  1.135e+01   8.194 2.84e-16 ***
## PARITY_truncated3               1.199e+02  1.323e+01   9.063  < 2e-16 ***
## PARITY_truncated4               1.232e+02  1.623e+01   7.592 3.44e-14 ***
## PARITY_truncated5+              1.085e+02  1.646e+01   6.591 4.58e-11 ***
## PLUR_truncated2                -3.251e+02  2.531e+01 -12.846  < 2e-16 ***
## PLUR_truncated3+               -4.121e+02  1.247e+02  -3.306 0.000951 ***
## smoking_typebefore and during -1.980e+02  1.518e+01 -13.041  < 2e-16 ***
```

```
## smoking_typebefore only       -3.233e+01  2.286e+01  -1.414 0.157250
## smoking_typeduring only       -1.353e+02  7.991e+01  -1.693 0.090497 .
## MAGE                           3.960e+00  8.084e-01   4.899 9.79e-07 ***
## MRACER1                        6.746e+01  1.409e+01   4.789 1.70e-06 ***
## MRACER2                       -1.044e+02  1.558e+01  -6.705 2.12e-11 ***
## MRACER3                        4.841e+01  4.026e+01   1.202 0.229216
## MRACER4                       -1.090e+02  7.363e+01  -1.480 0.138852
## MRACER5                       -2.978e+02  1.522e+02  -1.957 0.050401 .
## MRACER6                        1.935e+02  3.034e+02   0.638 0.523512
## MRACER7                        1.764e+02  7.941e+01   2.222 0.026328 *
## MRACER8                       -1.371e+02  2.759e+01  -4.971 6.79e-07 ***
## mortality                      1.656e+00  1.295e+00   1.279 0.200768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 428.5 on 9815 degrees of freedom
## Multiple R-squared:  0.5284, Adjusted R-squared:  0.5273
## F-statistic: 478.1 on 23 and 9815 DF,  p-value: < 2.2e-16
```

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




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

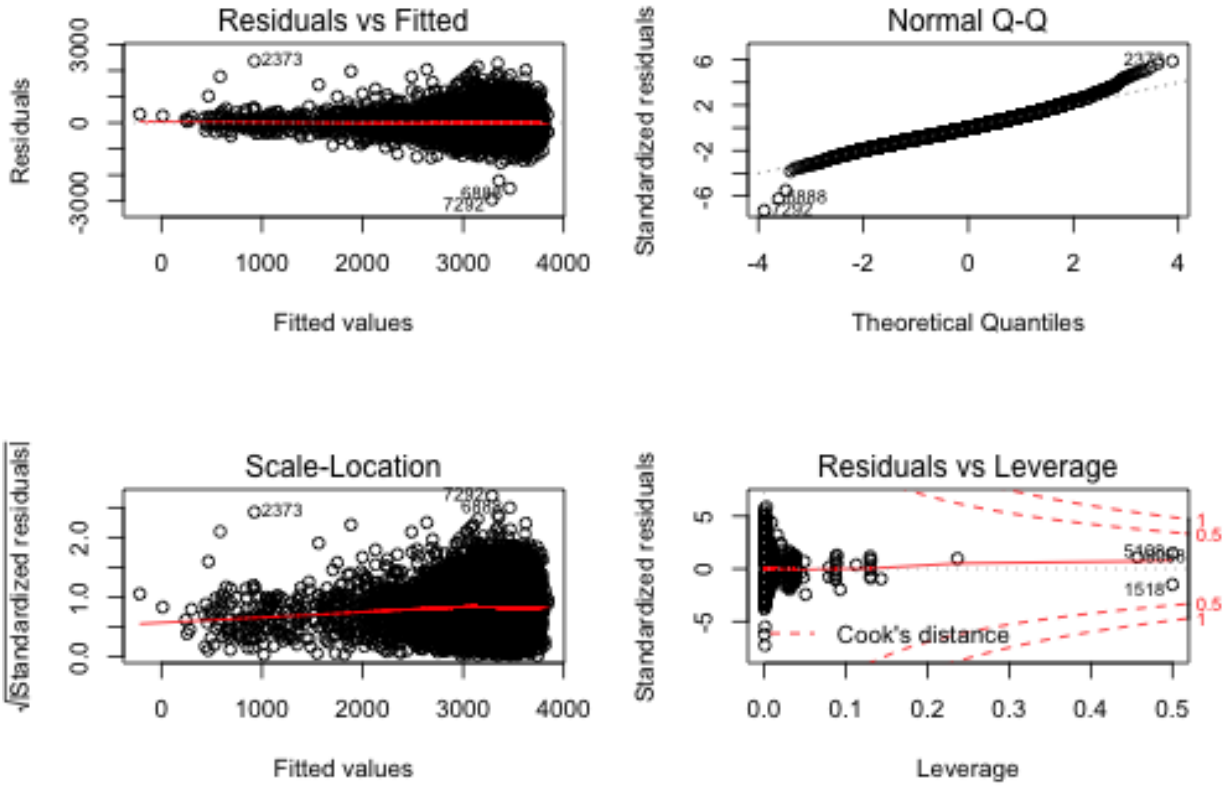
Robust on Model 4

```
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
## -2935.657  -272.922    -9.572   270.309  2369.618
##
## Coefficients:
##                               Value       Std. Error  t value
## (Intercept)                   -16081.5714   5756.7827     -2.7935
## GEST                            2541.9416    754.4087      3.3694
## GEST2                           -147.3543     36.4095     -4.0471
```

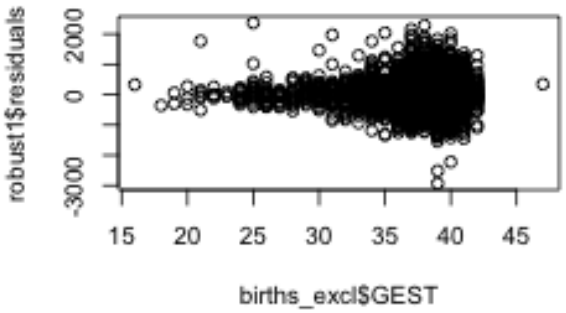
```
## GEST3                                3.7362      0.7682      4.8633
## GEST4                               -0.0334      0.0060     -5.5834
## PARITY_truncated2                   91.1205     11.0231      8.2663
## PARITY_truncated3                  120.8500     12.8445      9.4087
## PARITY_truncated4                  128.5594     15.7521      8.1614
## PARITY_truncated5+                 114.1579     15.9815      7.1431
## PLUR_truncated2                   -301.2572     24.5695    -12.2614
## PLUR_truncated3+                  -354.0980    121.0404     -2.9255
## smoking_typebefore and during     -202.0670     14.7393    -13.7094
## smoking_typebefore only           -16.9721      22.1926     -0.7648
## smoking_typeduring only           -142.2274     77.5799     -1.8333
## MAGE                                3.2846      0.7848      4.1850
## MRACER1                            80.1642     13.6747      5.8622
## MRACER2                          -100.9554     15.1209     -6.6765
## MRACER3                            36.4357     39.0884      0.9321
## MRACER4                           -88.2791     71.4787     -1.2350
## MRACER5                          -245.3286    147.7677     -1.6602
## MRACER6                           216.4640    294.5338      0.7349
## MRACER7                           181.1717     77.0908      2.3501
## MRACER8                          -119.8266     26.7875     -4.4732
## mortality                           1.3637      1.2568      1.0851
##
## Residual standard error: 403.6 on 9815 degrees of freedom
```

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



```r
# plot(robust1$fitted.values, robust1$residuals)
plot(births_excl$GEST, robust1$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
## 7292  352   39 -2935.65744 0.1849350
## 6888  945   39 -2519.80445 0.2154559
## 2373 3295   25  2369.61770 0.2291131
## 3958 5630   38  2285.69361 0.2375248
## 4759 1134   40 -2222.50329 0.2442787
## 9832 3232   38   116.11499 1.0000000
## 9833 3260   39   -43.19426 1.0000000
## 9834 2948   39  -422.22018 1.0000000
## 9835 3317   39   -66.35844 1.0000000
## 9836 2778   37  -378.71951 1.0000000
## 9837 3657   40    56.77502 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
}

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

#Train and test model2
model2_test_mse<-c()
for(i in 1:10){
  model2_train<-lm(data=train_list[[i]],BWTG~GEST+GEST2+PARITY_truncated+PLUR_truncated+smoking_type+MA
  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, mean(model2_test_mse))

#Train and test model3
model3_test_mse<-c()
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, mean(model3_test_mse))

#Train and test model4
model4_test_mse<-c()
for(i in 1:10){
  model4_train<-lm(data=train_list[[i]],BWTG~GEST+GEST2+GEST3+GEST4+PARITY_truncated+PLUR_truncated+smo
  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, mean(model4_test_mse))
```

```
robust1_test_mse<-c()
for(i in 1:10){
  robust1_train<-rlm(data=train_list[[i]],BWTG~GEST+GEST2+GEST3+PARITY_truncated+PLUR_truncated+smoking
  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, mean(robust1_test_mse))

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

As illustrated by the table above, model4 has the lowest MSE out of all models utilised. This implies that However, model3 is the second lowest and therefore exhibits the least amount of overfitting for a valid model.

```
#medqr <- rq(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_truncate
#summary(medqr)

#lowqr <- rq(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_truncate
#summary(lowqr)

#highqr <- rq(data = births_excl, BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_truncat
#summary(highqr)
```