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

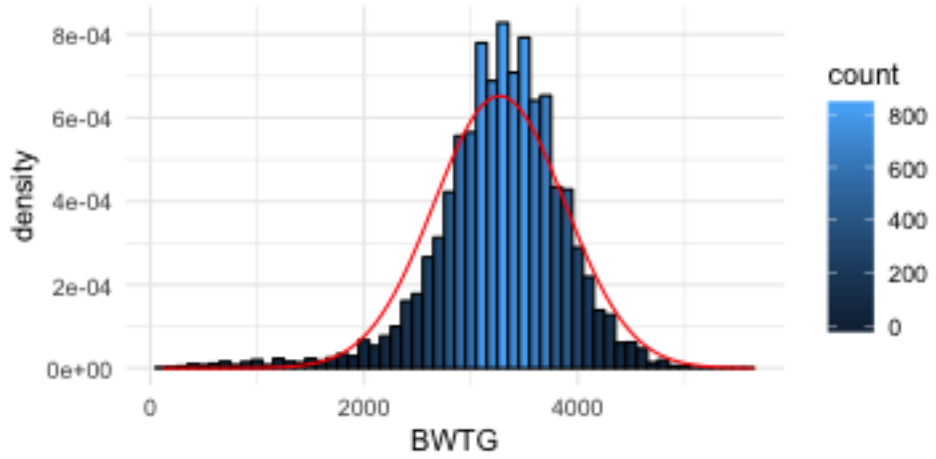
```
## read in data
births = read.csv("data/Yr1116Birth.csv", na.strings = "9999")
deaths = read.csv("data/Yr1116Death.csv")

## 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)
```

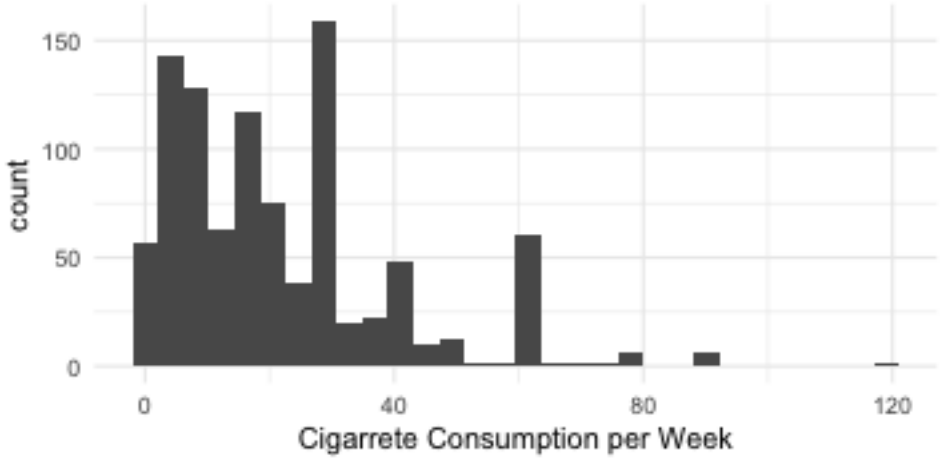
Exploratory Data Analysis

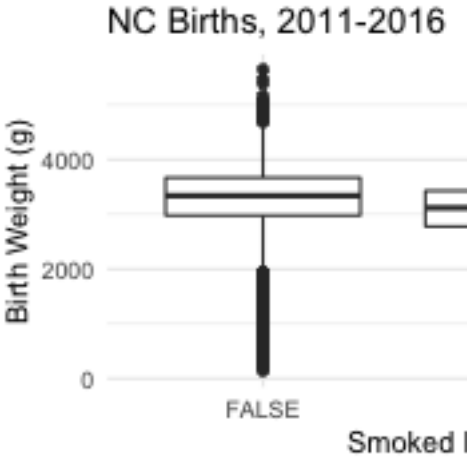
Birthweight

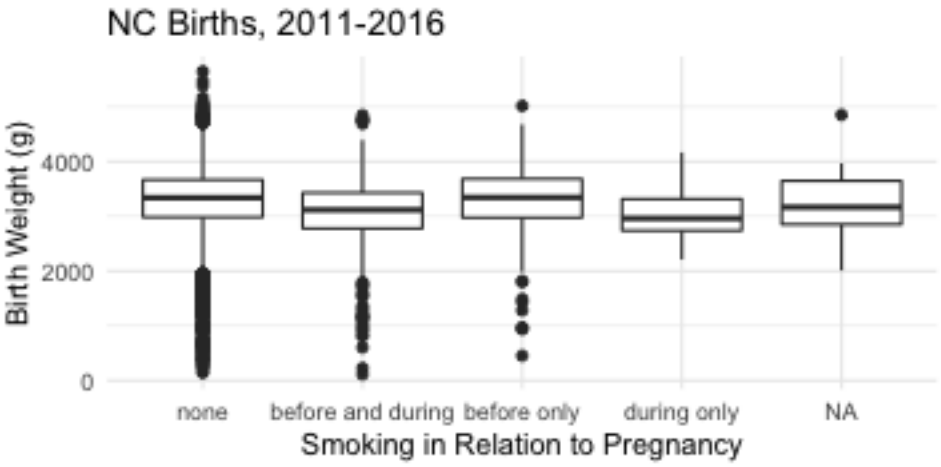


Birthweight is close to normally distrubted, with a slight left skew, centered around ~3300g with a standard deviation of 600g. There appear to be no large outliers in terms of birthweight. 430 birth weights are missing. We see that the left tail is much larger than we would expect in a normal distribution

Smoking

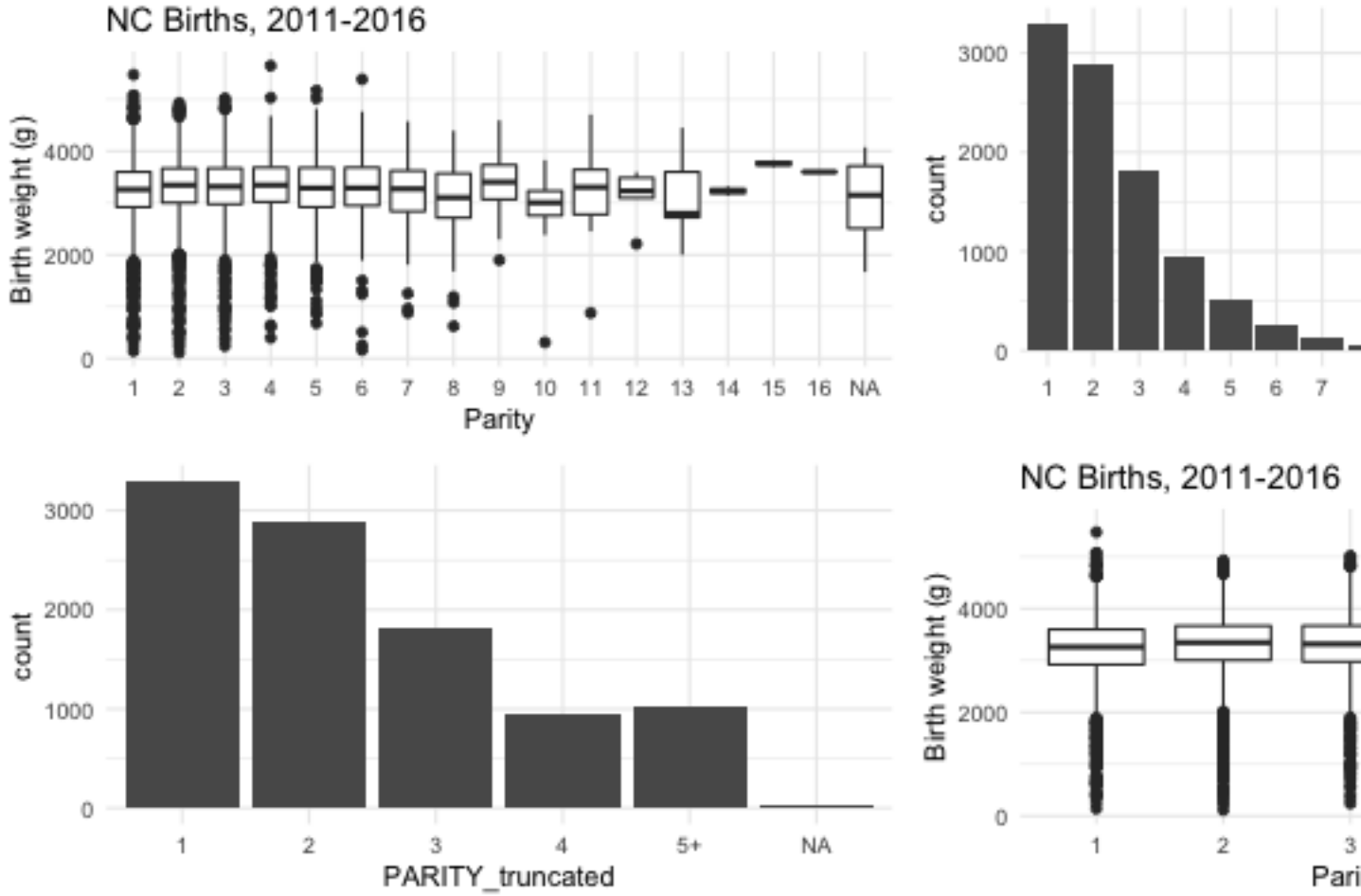






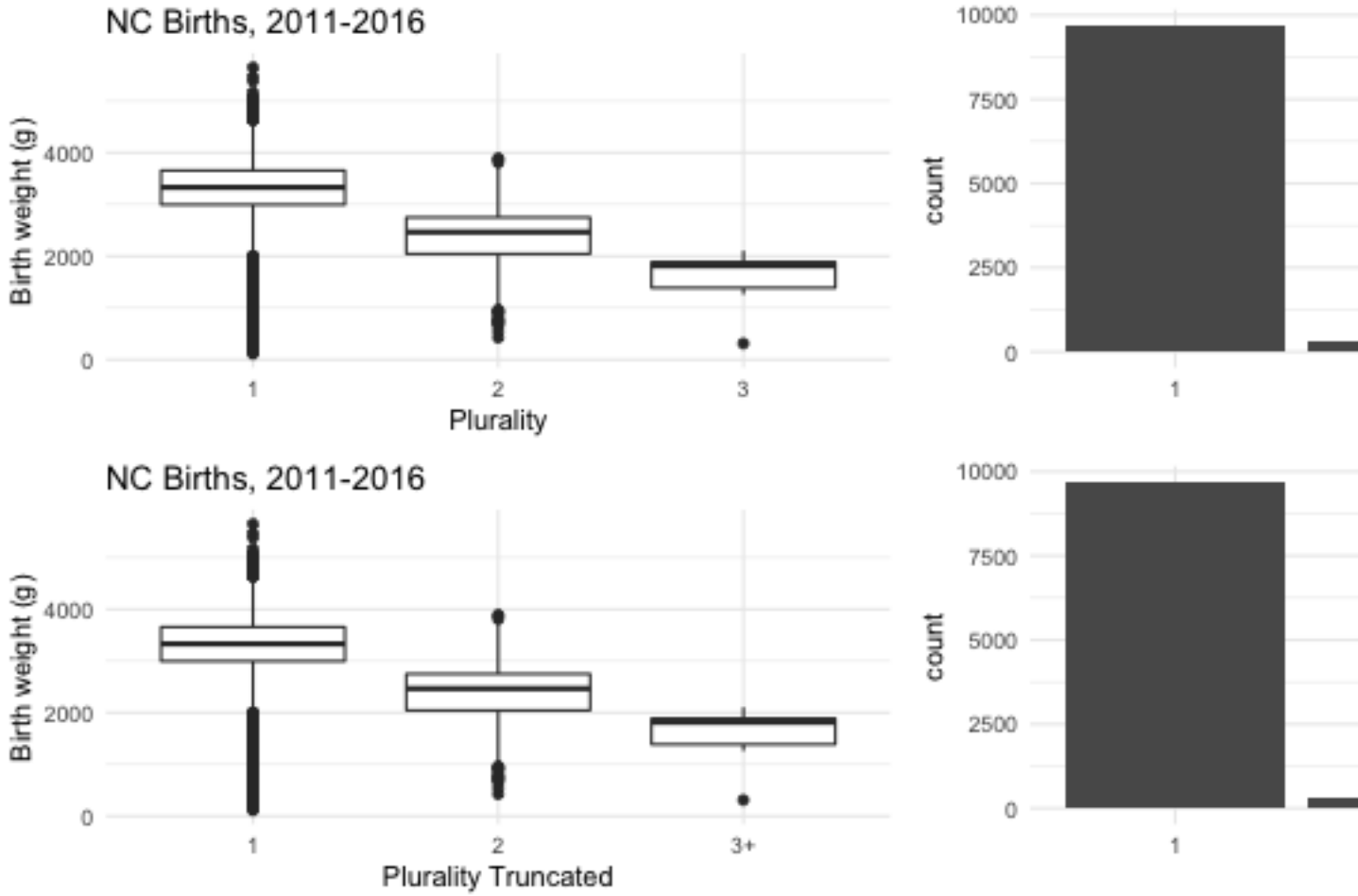
Around 13% of women smoked in the three months leading up to pregnancy and around 10% of women at any point during their pregnancy. Among those who did smoke during pregnance, the average number of cigarettes smoked during pregnancy was 23. The birthweight of children of smokers was significantly lower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



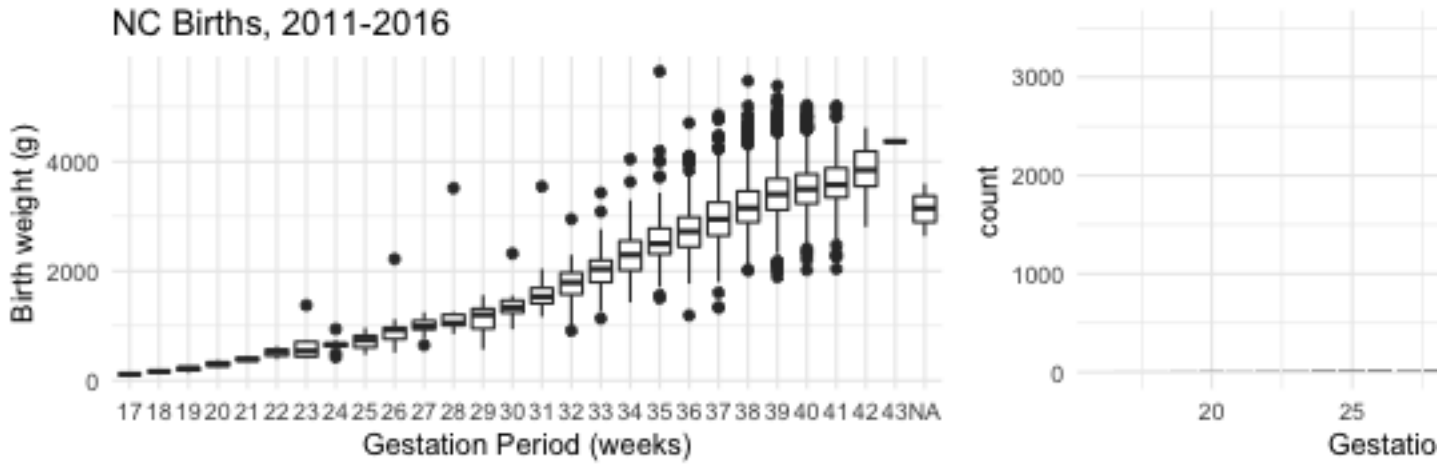
Independent of other variables, we see a negative relationship between parity and birth weight past the first child. The frequency of parity decreases in an exponential fashion. A second variable was created that truncates parities of at least five to improve interprability and prevent overfitting. The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

Plurality



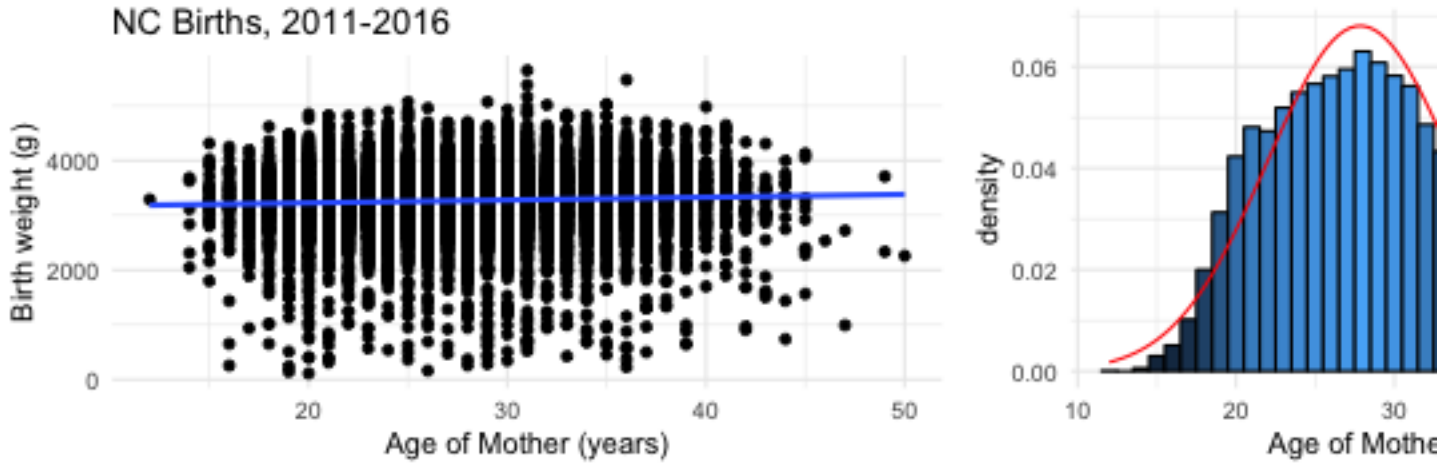
We see a strong non linear negative relationship between plurality and birth weight. The frequency of pluralities above two is extremely small, and we again see a proportionally small amount of missing data. A second variable was created that truncates pluralities of at least three to improve interprability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



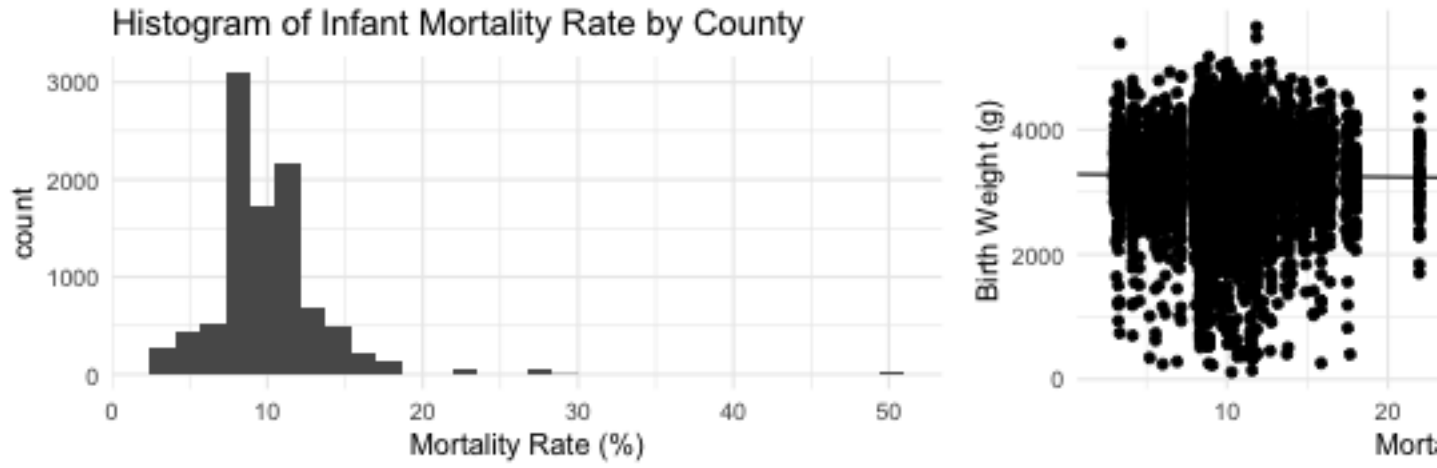
Mother's age seems to be fairly normally distributed with a mean of 27.7. There appears to be a positive relationship between the age of the mother and the birth weight. There is no evidence to suggest that the birth weight variance is not constant across the mother's age.

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



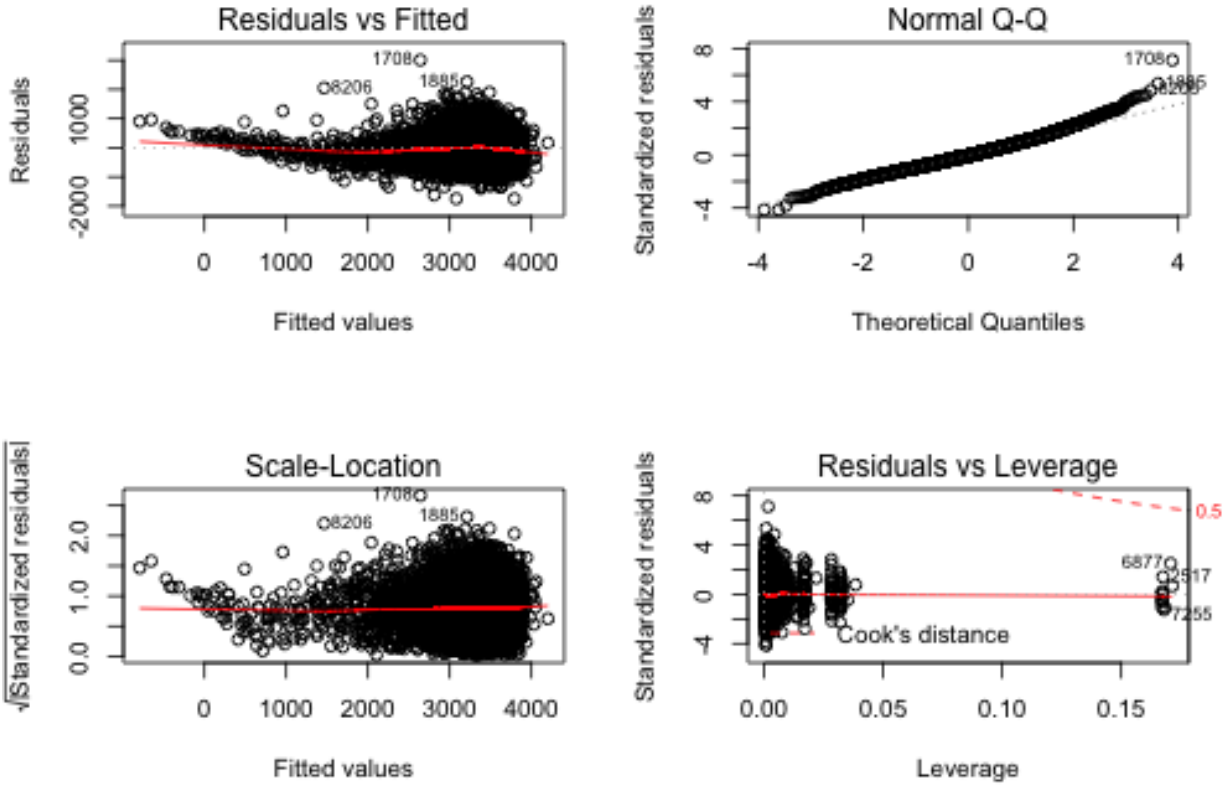
We chose to use infant mortality rate of birth county as a proxy for socioeconomic status, calculated as number of deaths before the age of 1 divided by total number of births in a county. The median county in the data had a infant mortality rate of 0.7%, with the range of infant mortality rates in our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

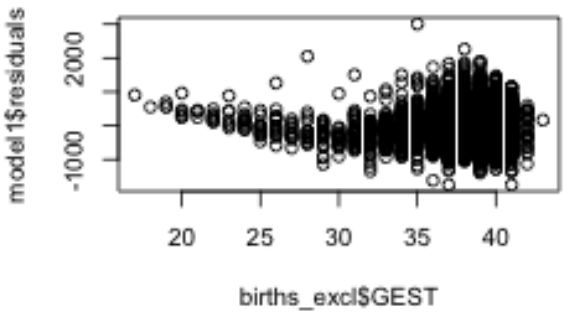
```
births_excl = na.omit(births)
births_excl = births_excl[which(births_excl$GEST < 80), ]
births_excl = births_excl %>%
  mutate(GEST2 = GEST^2, GEST3 = GEST ^ 3, GEST4 = GEST^4)
model1 = lm(data = births_excl, BWTG ~ GEST + PARITY_truncated + PLUR_truncated + smoking_type + MAGE +
summary(model1)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

```
##
## 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
## -1749.90  -286.55   -25.14   259.77  3007.27
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3826.6614    81.7933 -46.785  < 2e-16 ***
## GEST                           179.4493     1.9688  91.146  < 2e-16 ***
## PARITY_truncated2               91.6890    11.0112   8.327  < 2e-16 ***
## PARITY_truncated3              117.5208    12.9559   9.071  < 2e-16 ***
## PARITY_truncated4              150.9922    16.2409   9.297  < 2e-16 ***
## PARITY_truncated5+             126.7254    16.3932   7.730 1.18e-14 ***
## PLUR_truncated2               -356.9864    25.8515 -13.809  < 2e-16 ***
## PLUR_truncated3+              -542.3985   173.3440  -3.129  0.00176 **
## smoking_typebefore and during -188.5431    15.0363 -12.539  < 2e-16 ***
## smoking_typebefore only         22.9678    23.6484   0.971  0.33146
## smoking_typeduring only       -221.1307    75.0297  -2.947  0.00321 **
## MAGE                            3.2160     0.8115   3.963 7.45e-05 ***
## MRACER1                        83.2648    14.1032   5.904 3.67e-09 ***
## MRACER2                       -83.6466    15.4457  -5.416 6.25e-08 ***
## MRACER3                        22.7795    37.5783   0.606  0.54441
## MRACER4                        32.6577    55.4217   0.589  0.55570
## MRACER5                      -204.9559   173.4520  -1.182  0.23738
## MRACER7                        19.8608    71.7657   0.277  0.78198
## MRACER8                       -81.1929    26.9967  -3.008  0.00264 **
## mortality                       1.6592     1.1823   1.403  0.16053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 423.1 on 9851 degrees of freedom
## Multiple R-squared:  0.5223, Adjusted R-squared:  0.5214
## F-statistic: 566.9 on 19 and 9851 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.

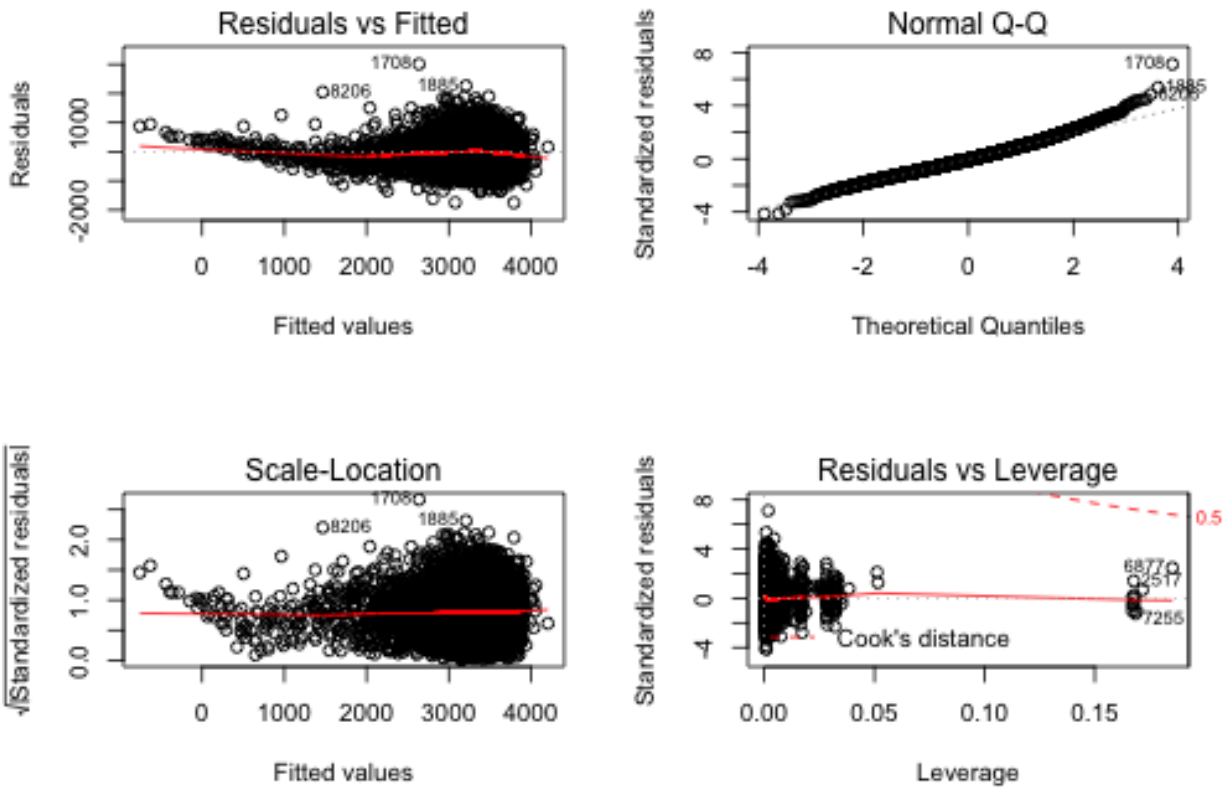
The residual graph for Mother's race indicates that residuals are lower for for races 3, 4, 5, 6, and 7 and higher for the other races. This could be something to explore.

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).

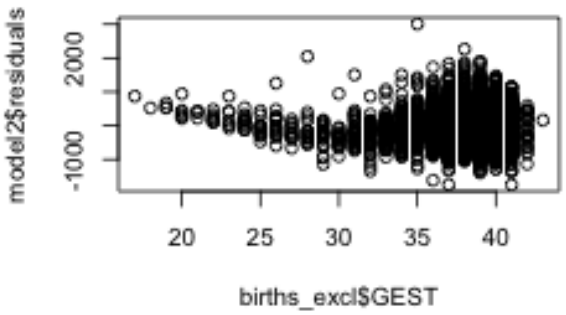
```
model2 = lm(data = births_excl, BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated + smoking_type +
summary(model2)
```

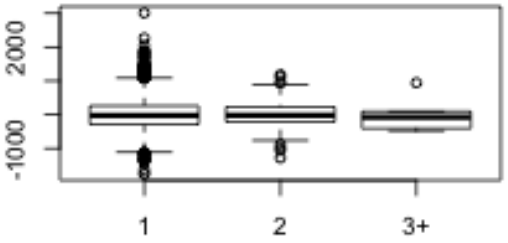
```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + PARITY_truncated + PLUR_truncated +
##     smoking_type + MAGE + MRACER + mortality, data = births_excl,
##     na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1752.12  -286.46   -25.11   259.47  3009.35
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3693.9365   346.9284 -10.648  < 2e-16 ***
## GEST                           171.6335    19.9511   8.603  < 2e-16 ***
## GEST2                            0.1128     0.2866   0.394  0.69383
## PARITY_truncated2               91.8994    11.0246   8.336  < 2e-16 ***
## PARITY_truncated3              117.8758    12.9878   9.076  < 2e-16 ***
## PARITY_truncated4              151.4200    16.2779   9.302  < 2e-16 ***
## PARITY_truncated5+             127.1083    16.4227   7.740 1.09e-14 ***
## PLUR_truncated2               -355.9064    25.9978 -13.690  < 2e-16 ***
## PLUR_truncated3+              -544.3576   173.4228  -3.139  0.00170 **
## smoking_typebefore and during -188.4195    15.0402 -12.528  < 2e-16 ***
## smoking_typebefore only         22.9072    23.6499   0.969  0.33277
## smoking_typeduring only       -220.8929    75.0353  -2.944  0.00325 **
## MAGE                            3.2181     0.8115   3.966 7.37e-05 ***
## MRACER1                        83.3669    14.1062   5.910 3.54e-09 ***
## MRACER2                       -83.5207    15.4497  -5.406 6.60e-08 ***
## MRACER3                        23.0625    37.5868   0.614  0.53951
## MRACER4                        32.9520    55.4292   0.594  0.55220
## MRACER5                      -205.4416   173.4638  -1.184  0.23630
## MRACER7                        20.2263    71.7748   0.282  0.77810
## MRACER8                       -80.9032    27.0079  -2.996  0.00275 **
## mortality                       1.6508     1.1825   1.396  0.16274
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 423.1 on 9850 degrees of freedom
## Multiple R-squared:  0.5223, Adjusted R-squared:  0.5214
## F-statistic: 538.5 on 20 and 9850 DF,  p-value: < 2.2e-16
```

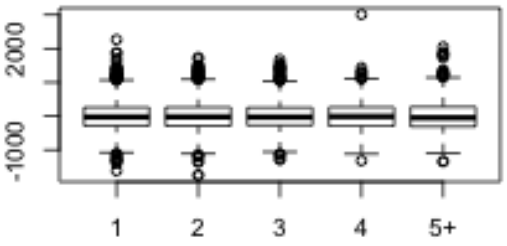
```
par(mfrow = c(2,2))
plot(model2)
```

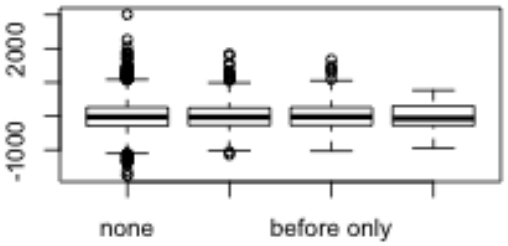



```
# plot(model2$fitted.values, model2$residuals)
plot(births_excl$GEST, model2$residuals)
plot(births_excl$PLUR_truncated, model2$residuals)
plot(births_excl$PARITY_truncated, model2$residuals)
plot(births_excl$smoking_type, model2$residuals)
```

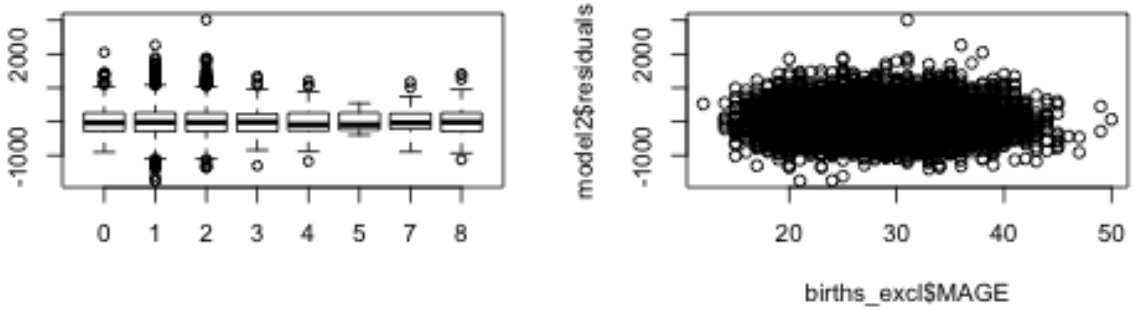









```
plot(births_excl$MRACER, model2$residuals)
plot(births_excl$MAGE, model2$residuals)
```



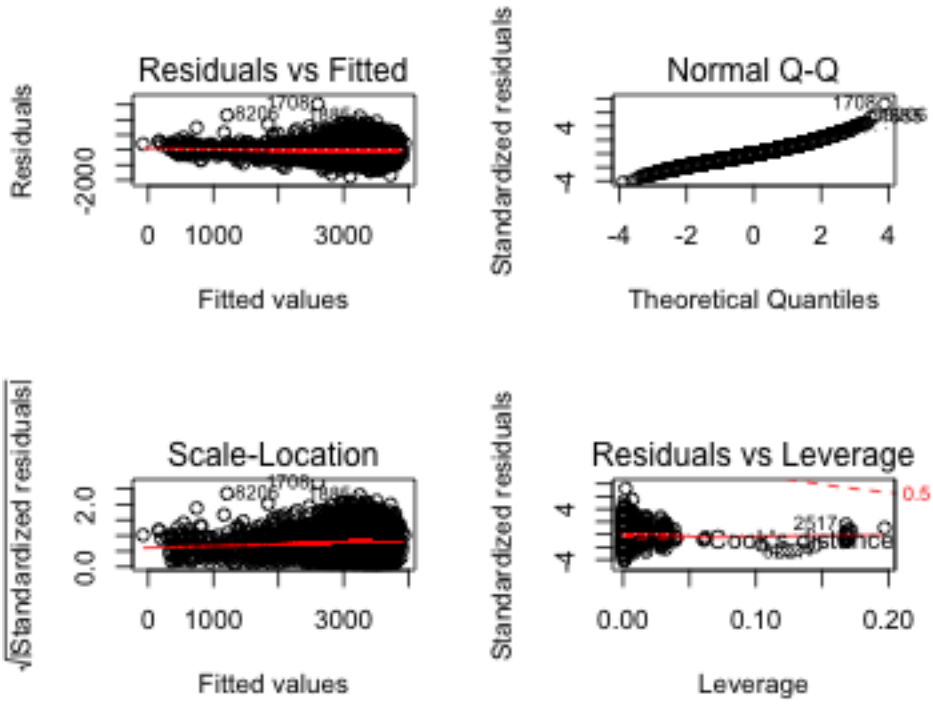
The new model still displays the original downwards trend in the residual vs gestational period graph. Perhaps another transformation on GEST would be helpful – a cubic term can be added. The other residuals plots also retain their trends from model 1.

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

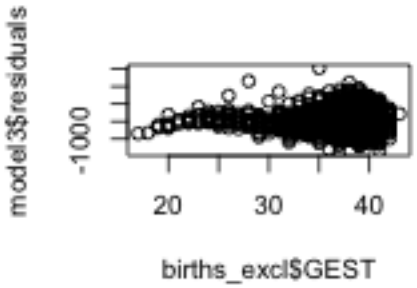
```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + GEST3 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1767.75  -282.62   -23.31   255.78  3045.09
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.564e+04  1.625e+03   9.626  < 2e-16 ***
## GEST                        -1.684e+03  1.537e+02 -10.958  < 2e-16 ***
## GEST2                        5.774e+01  4.742e+00  12.178  < 2e-16 ***
## GEST3                       -5.829e-01  4.788e-02 -12.176  < 2e-16 ***
## PARITY_truncated2            8.402e+01  1.096e+01   7.664 1.97e-14 ***
## PARITY_truncated3            1.075e+02  1.292e+01   8.317  < 2e-16 ***
## PARITY_truncated4            1.404e+02  1.618e+01   8.675  < 2e-16 ***
```

```
## PARITY_truncated5+             1.178e+02  1.632e+01   7.219 5.62e-13 ***
## PLUR_truncated2               -3.273e+02  2.591e+01 -12.631  < 2e-16 ***
## PLUR_truncated3+              -5.814e+02  1.722e+02  -3.377 0.000736 ***
## smoking_typebefore and during -1.890e+02  1.493e+01 -12.658  < 2e-16 ***
## smoking_typebefore only        2.583e+01  2.348e+01   1.100 0.271203
## smoking_typeduring only       -2.238e+02  7.448e+01  -3.005 0.002667 **
## MAGE                           3.352e+00  8.056e-01   4.161 3.20e-05 ***
## MRACER1                        8.125e+01  1.400e+01   5.803 6.73e-09 ***
## MRACER2                       -8.658e+01  1.534e+01  -5.645 1.70e-08 ***
## MRACER3                        2.335e+01  3.731e+01   0.626 0.531428
## MRACER4                        2.127e+01  5.503e+01   0.387 0.699129
## MRACER5                       -1.933e+02  1.722e+02  -1.123 0.261526
## MRACER7                        2.415e+01  7.124e+01   0.339 0.734658
## MRACER8                       -9.154e+01  2.682e+01  -3.413 0.000646 ***
## mortality                      1.372e+00  1.174e+00   1.168 0.242702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 420 on 9849 degrees of freedom
## Multiple R-squared:  0.5294, Adjusted R-squared:  0.5284
## F-statistic: 527.6 on 21 and 9849 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)
```



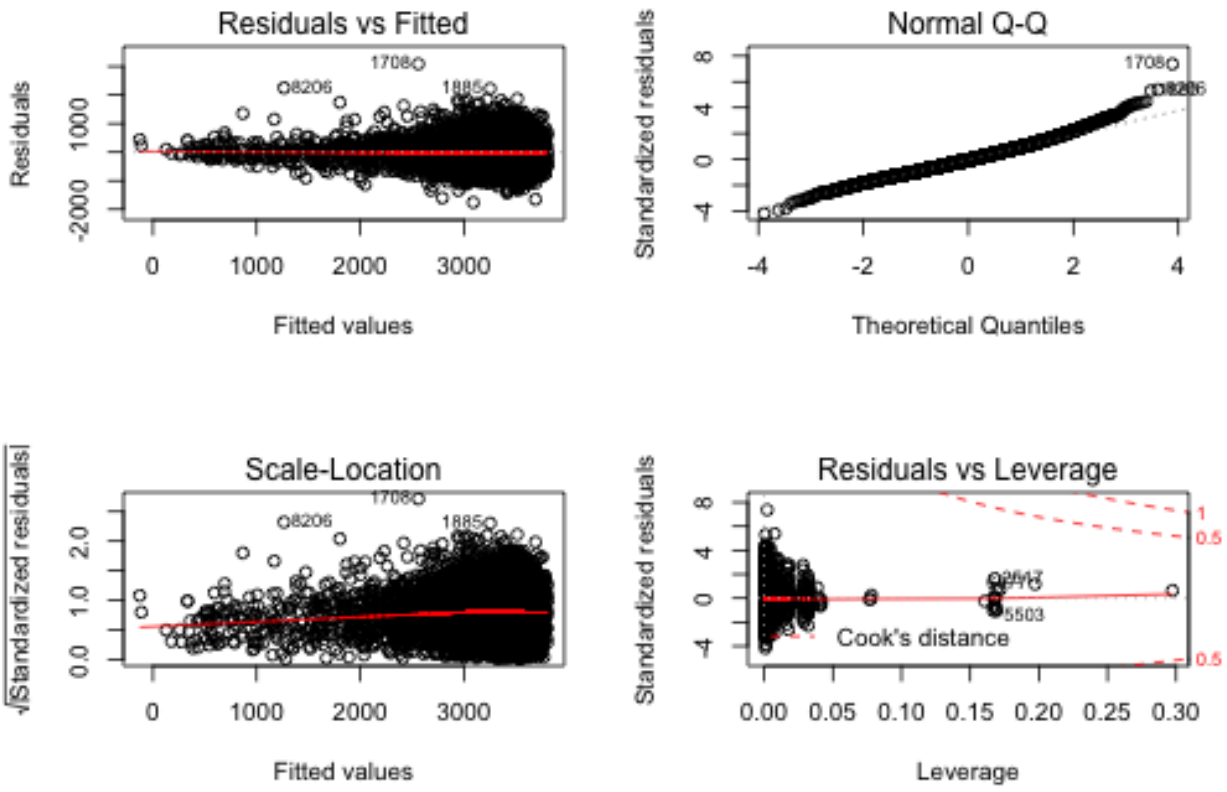
The residual vs gestational period shows a much more random pattern than before. It is worth investigating if adding a quartic term would help. The other residuals plots also retain their trends from model 1.

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

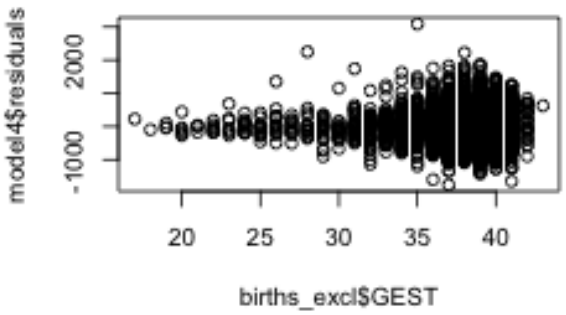
```
##
## Call:
## lm(formula = BWTG ~ GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality,
##     data = births_excl, na.action = "na.exclude")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1758.85  -279.95   -23.01   255.71  3087.12
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -2.225e+04  6.975e+03  -3.190 0.001429 **
## GEST                            3.377e+03  9.191e+02   3.674 0.000240 ***
## GEST2                          -1.894e+02  4.451e+01  -4.256 2.10e-05 ***
## GEST3                           4.666e+00  9.411e-01   4.958 7.23e-07 ***
## GEST4                          -4.103e-02  7.346e-03  -5.585 2.40e-08 ***
## PARITY_truncated2               8.231e+01  1.095e+01   7.517 6.10e-14 ***
## PARITY_truncated3               1.051e+02  1.291e+01   8.143 4.31e-16 ***
## PARITY_truncated4               1.379e+02  1.616e+01   8.529  < 2e-16 ***
## PARITY_truncated5+              1.171e+02  1.629e+01   7.189 7.03e-13 ***
## PLUR_truncated2                -3.158e+02  2.595e+01 -12.167  < 2e-16 ***
## PLUR_truncated3+               -5.265e+02  1.722e+02  -3.058 0.002236 **
## smoking_typebefore and during -1.869e+02  1.491e+01 -12.537  < 2e-16 ***
```

```
## smoking_typebefore only        2.673e+01  2.344e+01   1.140 0.254233
## smoking_typeduring only       -2.208e+02  7.437e+01  -2.969 0.002991 **
## MAGE                           3.310e+00  8.044e-01   4.115 3.90e-05 ***
## MRACER1                        8.017e+01  1.398e+01   5.733 1.02e-08 ***
## MRACER2                       -8.737e+01  1.531e+01  -5.705 1.20e-08 ***
## MRACER3                        2.410e+01  3.725e+01   0.647 0.517688
## MRACER4                        1.766e+01  5.495e+01   0.321 0.747856
## MRACER5                       -1.867e+02  1.719e+02  -1.086 0.277612
## MRACER7                        2.399e+01  7.114e+01   0.337 0.735931
## MRACER8                       -9.485e+01  2.679e+01  -3.541 0.000401 ***
## mortality                      1.564e+00  1.173e+00   1.333 0.182449
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 419.3 on 9848 degrees of freedom
## Multiple R-squared:  0.5309, Adjusted R-squared:  0.5298
## F-statistic: 506.6 on 22 and 9848 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.

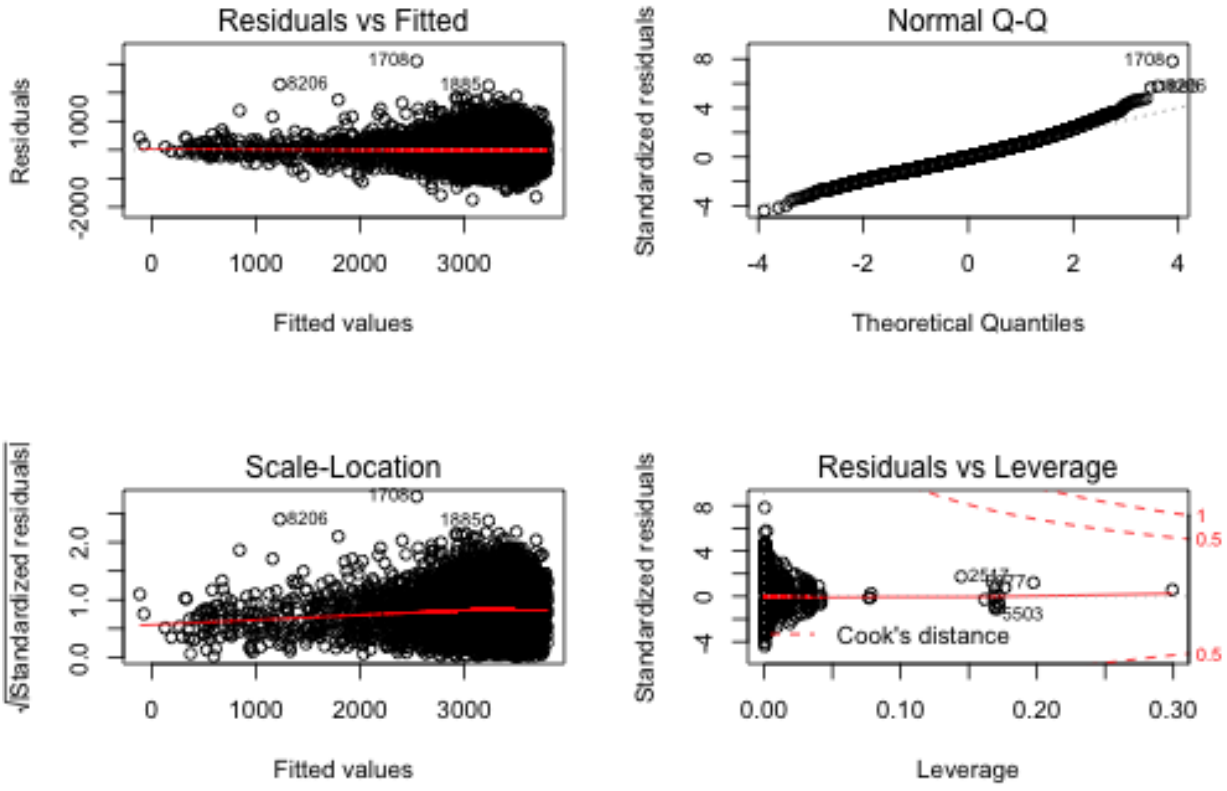
Robust on Model 4

```
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
## -1746.017  -267.642    -9.236   268.502  3106.390
##
## Coefficients:
##                               Value       Std. Error  t value
## (Intercept)                   -20776.8407   6822.8388     -3.0452
## GEST                            3186.2379    899.1080      3.5438
## GEST2                           -180.3156     43.5416     -4.1412
```

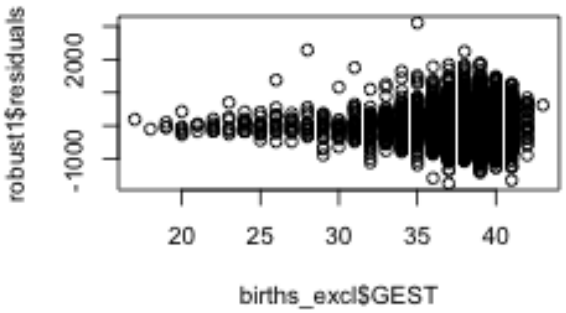
```
## GEST3                              4.4737     0.9206     4.8594
## GEST4                             -0.0395     0.0072    -5.4975
## PARITY_truncated2                 84.0226    10.7116     7.8441
## PARITY_truncated3                109.5493    12.6265     8.6762
## PARITY_truncated4                142.7045    15.8128     9.0246
## PARITY_truncated5+               113.8826    15.9402     7.1444
## PLUR_truncated2                 -297.5779    25.3900   -11.7203
## PLUR_truncated3+                -509.3789   168.4389    -3.0241
## smoking_typebefore and during   -188.6272    14.5863   -12.9318
## smoking_typebefore only           20.6486    22.9310     0.9005
## smoking_typeduring only         -206.9120    72.7510    -2.8441
## MAGE                              2.7993     0.7869     3.5575
## MRACER1                          85.9511    13.6787     6.2836
## MRACER2                         -85.6589    14.9815    -5.7176
## MRACER3                          18.4970    36.4417     0.5076
## MRACER4                           9.5974    53.7520     0.1785
## MRACER5                        -187.4004   168.1848    -1.1143
## MRACER7                          22.9138    69.5883     0.3293
## MRACER8                         -89.9993    26.2052    -3.4344
## mortality                         1.3815     1.1472     1.2042
##
## Residual standard error: 397.4 on 9848 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
## 1708 5648   35 3106.3903 0.1720773
## 8206 3515   28 2285.8996 0.2338419
## 1885 5475   38 2241.4920 0.2384755
## 8147 5386   39 1887.2395 0.2832394
## 9830 5075   39 1865.6214 0.2865212
## 9865 3619   40  120.9297 1.0000000
## 9866 3147   41 -490.9088 1.0000000
## 9868 3798   40  204.7097 1.0000000
## 9869 2745   36  -51.6868 1.0000000
## 9870 3856   40  354.2595 1.0000000
## 9871 3147   39 -407.5821 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))
```

```r
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.

```
{r} #  # medqr <- rq(data = births_excl, BWTG ~ GEST + GEST2 +
GEST3 + GEST4 + PARITY_truncated + PLUR_truncated + smoking_type
+ MAGE + MRACER + mortality, na.action = "na.exclude") # summary(medqr)
#  # lowqr <- rq(data = births_excl, BWTG ~ GEST + GEST2 +
GEST3 + GEST4 + PARITY_truncated + PLUR_truncated + smoking_type
+ MAGE + MRACER + mortality, na.action = "na.exclude", tau=0.05)
# summary(lowqr) #  # highqr <- rq(data = births_excl, BWTG ~
GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + PLUR_truncated
+ smoking_type + MAGE + MRACER + mortality, na.action = "na.exclude",
tau=0.95) # summary(highqr) #
```