**440 Case Study 1**

# *Group 8: Jake Epstein, Daniel Spottiswood, Michael Tan, Sahil Patel, Man-Lin Hsiao 9/17/2019*

## Introduction

Birth weight is a commonly-used indicator of a newborn infant’s health status, tracked closely by the World Health Organization and governments.

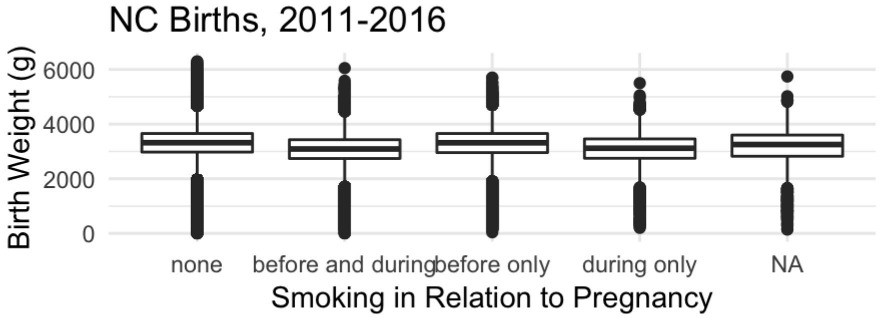
The goals of this case study are as follows: 1) determining which factors are associated with birth weight 2) modeling the relationship between each factor and birth weight 3) exploring whether fit of linear regression model is adequate or whether robust regression methods are needed 4) estimating the impact of eliminating maternal smoking on birth weight

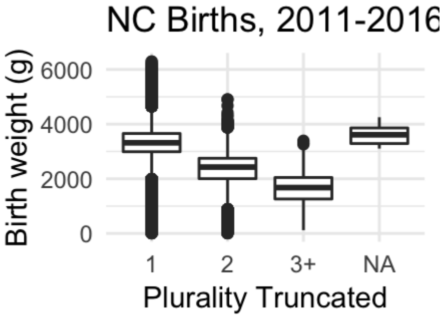
## Data Cleaning & Exploratory Data Analysis

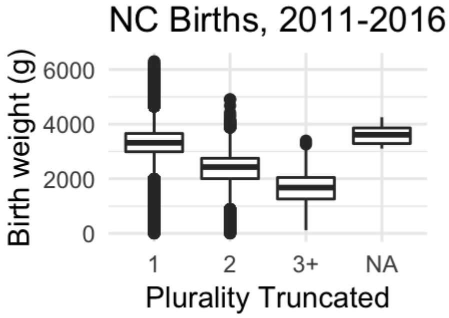
For data cleaning and processing, we chose to: (1) make smoking a categorical variable, (2) truncate the variables plurality and (3) parity, (4) consolidate the race variable, and (5) calculate infant mortality for each county.

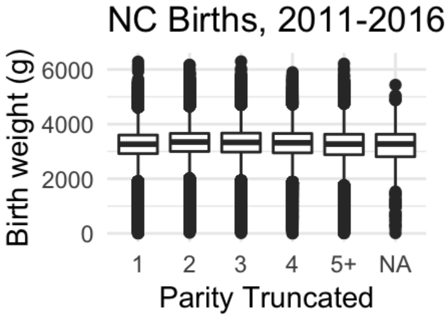
### Smoking

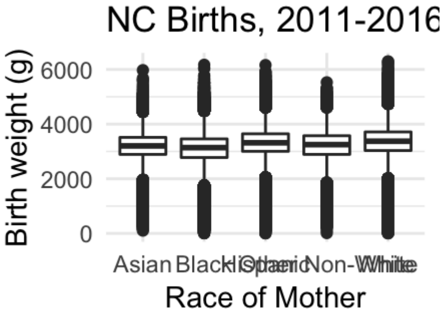
Independent of other variables, we see a negative relationship between parity and birth weight past the first child. The frequency of parity decreases in an exponential fashion. A second variable was created that truncates parities of at least five to improve interprability and prevent overfitting.











Around 13% of women smoked in the three months leading up to pregnancy and around 10% of women smoked at any point during their pregnancy. Among those who did smoke during pregnancy, the average number of cigarettes smoked during pregnancy was 23. The birthweight of children of mothers who smoked before and during pregnancy was significantly lower than that of the children of nonsmokers, with an average difference of about 200 grams. There is also a significant relationship between birthweight and smoking before pregnancy, even for those who did not smoke during pregnancy.

### 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 interpratability and prevent overfitting.

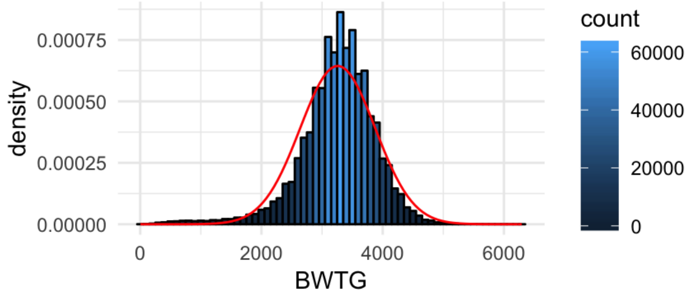
### Parity

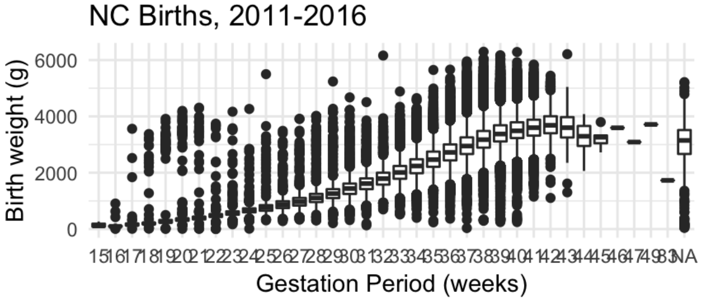
Independent of other variables, we see a negative relationship between parity and birth weight past the first child. The frequency of parity decreases in an exponential fashion. A second variable was created that truncates parities of at least five to improve interprability and prevent overfitting. The quantity of missing data is relatively small, as there are very few instances of mothers having more than 5 children.

### Race of Mother

There are significant differences between the average birth weights of mothers of different races. We see that mothers that self-identified as white have the largest mean baby weight at 3.34 kg, while black mothers have the lowest mean baby weight at only 3.07 kg. 56 percent of mothers identify as white, 24 percent identify as black, 15 percent identify as hispanic, and 4 percent identify as Asian.

### EDA on Birth Weight





Birthweight is close to normally distributed, with a slight left skew. There appears 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. We'll keep this in mind when evaluating the assumptions of linear regression.

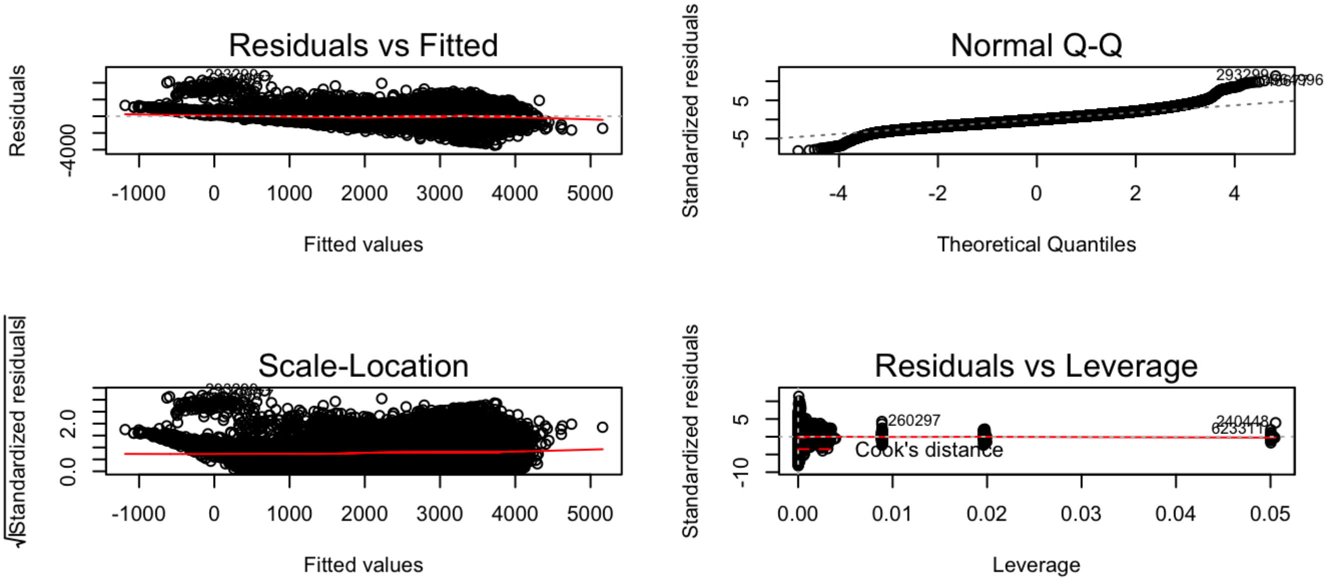
### EDA on Gestation

There appears to be a non-linear positive relationship between gestation period and birth weight. The distribution is left skewed. 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 it is biologically impossible for a human to gestate for 83 weeks, we will exclude it from our analysis when building the model.

## <u>Model Selection and Validation</u>

In building the model, we chose to drop the point at 80 weeks of gestation, which is likely an outlier that has no reason to be there, as no human can possibly gestate for 80 weeks (~1.54 years). Additionally, we dropped all rows with missing data.

### Model 1: Initial Approach



We noticed that the residuals vs fitted values and residuals vs gestational period plot slope downwards, indicating that there is a departure from linearity. A transformation may be helpful. The model may improve with additional higher-order terms added for gestation. Furthermore, the Q-Q plot indicates a departure from normality, likely as a result of the left skew of birthweight. We will keep this in mind as we make improvements to this initial model.

The residual graph for Smoking has higher residuals for no smoking than for smoking of any kind.

The GVIF outputs also do not indicate that there is multicollinearity.

### Interaction Effects

To explore interaction effects, we explored potential interactions that had logical bases in reality – meaning that there was evidence in our daily experiences and background knowledge to suggest that an interaction may exist. We explored interactions between 1) race and county infant mortality 2) race and smoking 3) race and parity and 4) race and age.

After running a regression exploring all those interactions listed above, we found that race and county infant mortality had significant interaction effects on birth weight.

### Higher Order Terms

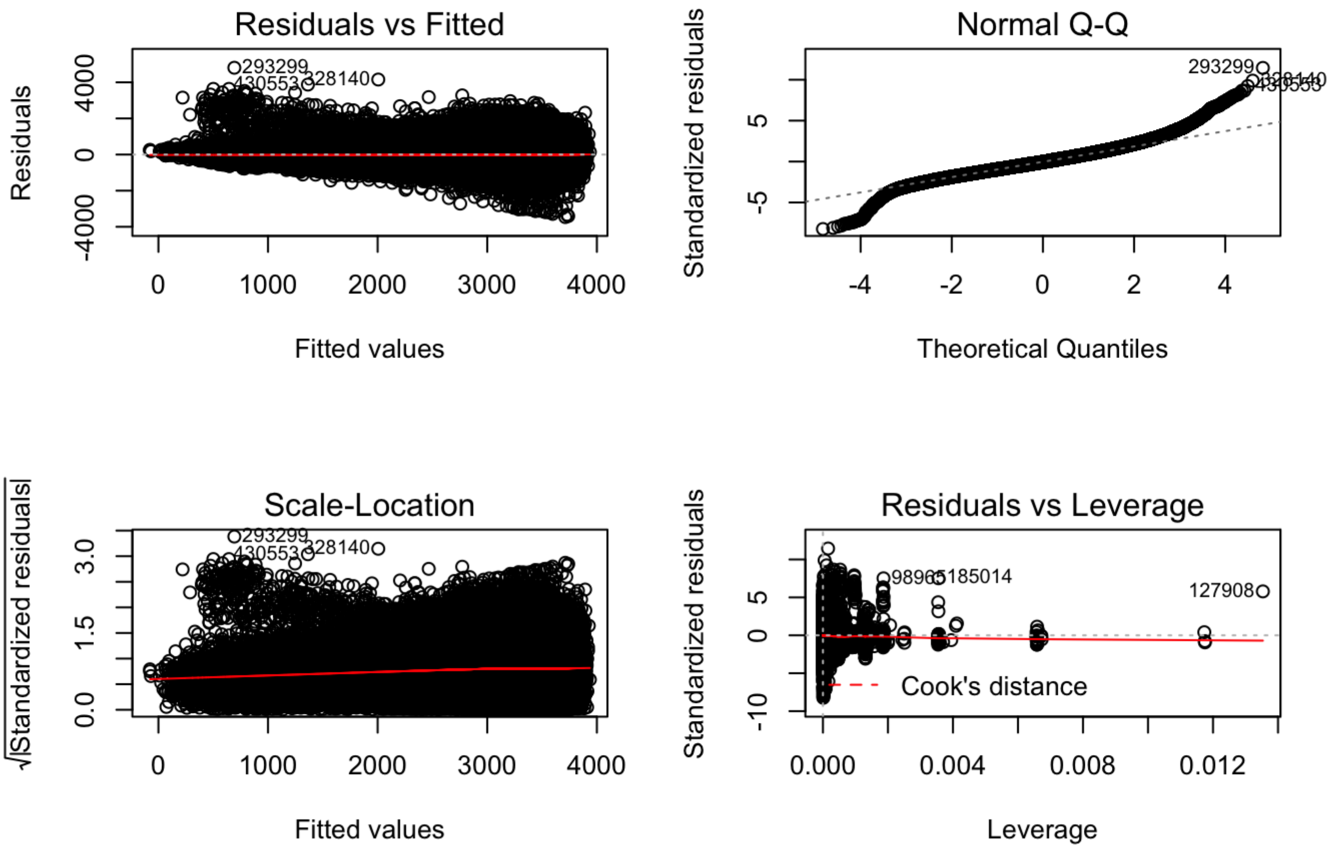
For next steps, we explored adding higher order terms for gestation onto our model.

Quadratic: While adding a quadratic term was statistically significant, the new model still displayed the original downwards trend in the residual vs gestational period graph. The other residuals plots also retain their trends from model 1.

Cubic: After adding a cubic term to gestation, the residual vs gestational period shows a much more random pattern than before and the cubic term is statistically significant. The other residuals plots also retain their trends from model 1.

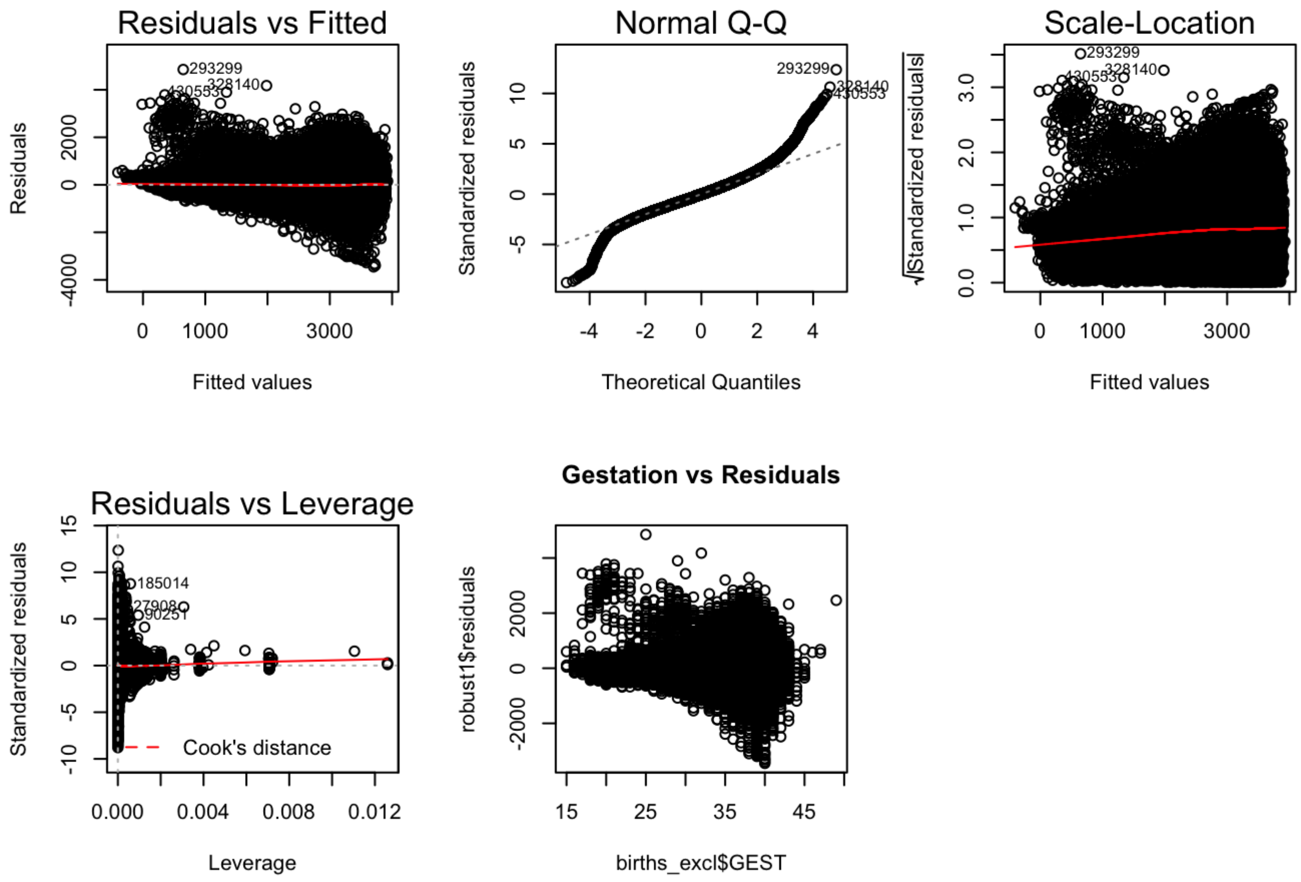
### Model 4: Adding Higher Order Terms to Gestation

```
## 
## Call:
## lm(formula = BWTG ~ SEX + GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated + 
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality + 
##     MRACER:mortality, data = births_excl, na.action = "na.exclude")
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3461.8  -276.4   -19.0   254.1  4807.4 
## 
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 -1.722e+04  7.563e+02  -22.768  < 2e-16
## SEXMale                      1.251e+02  9.917e-01  126.114  < 2e-16
## GEST                         2.771e+03  9.981e+01   27.762  < 2e-16
...
```



Quartic: The addition of the quartic term further smooths the residual plot, and it appears that additional polynomial degrees would only lead to overfitting and loss of interprability. Despite this, our assumption of normality still does not appear to be totally met, as shown by the Q-Q plot, and the residuals are not quite homoskedastic, as the variance of the residuals increases as the fitted values gets larger. Because of this, we will look into utilizing M-estimation.

### Robust on Model 4



```
##        bwt gest      resid    weight
## 293299 5500   25 4853.66504 0.1088509
## 328140 6160   32 4176.24242 0.1265090
## 430553 5239   29 3898.51301 0.1355205
## 564996 4139   20 3788.47364 0.1394564
## 48677  4281   21 3753.43242 0.1407578
## 717861 3289   39  -27.74374 1.0000000
## 717862 3770   41   67.88786 1.0000000
## 717863 2872   39 -318.78756 1.0000000
## 717864 3360   38  113.08005 1.0000000
## 717866 3713   40  126.52370 1.0000000
## 717867 3180   39 -356.63966 1.0000000
```

Our robust method downweights the weighting on observations with the largest residuals, as shown above. Looking at the output, it appears we still run into the same challenges as before with normality, even with this more robust approach. We will compare the approaches below by MSE using cross validation and splitting the data into test and training sets, and then we will use quantile regression to explore these relationships further.

## Cross Validation of Model

```
##               model1   model2   model3   model4  robust1
## Average MSE 180022.8 179896.5 176877.2 176349.0 176540.4
```

We employed K-fold cross-validation to test the models on their ability to accurately predict. Following the technique, we shuffled the data to randomize the sample, and then partitioned the data into K=10 folds (groups). We used the 10 folds to form 10 pairs of test and training sets; every test set is composed a differnet fold of the 10 total folds, and each training set is the remaining nine folds combined. We used the training sets to produce the estimates from the model, and use the test sets to produce an error score, the MSE. We averaged the 10 MSEs for all for 5 the models we developed and used this output as our metric of comparison.

We prefer using these MSE scores over the in-sample results reported in our initial model estimates. Using cross validation allows us to mitigate the adverse effects of overfitting our model by adding more covariates. As illustrated by the table above, model4 has the lowest MSE out of all models utilised. This implies that model4 yields the least erroneous out of sample (test data) results, and thus indicates the least amount of overfitting. Notably, the robust model had the second lowest MSE. The robust model reduces the influence of outlying data, therefore we will consider also solving for the MAE across our models as it may increase the interpretability of the robust error. This indicates that by attempting to reduce the influence of outlying data in our model, we may have begun to overfit the data.

## Quantile Regression

- Main focus: how does smoking effect vary based on birthweight?

```
##
## Call: rq(formula = BWTG ~ SEX + GEST + GEST2 + GEST3 + GEST4 + PARITY_truncated +
##     PLUR_truncated + smoking_type + MAGE + MRACER + mortality +
##     MRACER:mortality, tau = c(0.05, 0.25, 0.5, 0.75, 0.95))
...
```

```
##           Smoking: before Smoking: during Smoking: before & during
## LR Model4          -11.50         -115.30                  -203.40
## QR 0.05            -23.20         -193.40                  -200.90
## QR 0.25            -20.20         -122.10                  -207.80
## QR 0.5              -2.70         -104.50                  -198.10
## QR 0.75              5.80         -110.90                  -187.00
## QR 0.95             54.90          -45.93                  -169.50
```

## Interpretations and Conclusions

### Model 4

Since our Model 4 seemed to fit the data best, we'll report our conclusions using this model. Its output is replicated above.

Overall, the model does a good job of explaining birthweight, with our selected variables accounting for about 55% of the variation in birthweight in this sample, and each of the variables was highly predictive of birthweight.

**Sex** - On average male babies are heavier by about 125 grams, holding all else constant

**Gestation** - Given that there is a polynomial relationship between gestation and birthweight, it is hard to interpret this relationship simply, but in general, longer gestation periods are expected to yield heavier babies.

**Parity** - In general, children with a higher parity, that is to say children with more older siblings, tend to be heavier than first-borns. Second-borns are expected to be 85 grams heavier than the average first born, holding all else constant. Other higher-parity children are expected to be even heavier, by about 100 grams versus a first-born.

**Plurality** - Twins and triplets (or more) are expected to be much lighter than single-birth babies, by about 310-320 grams.

**Smoking** - smoking is expected to have a negative effect on babies' birthweights. Mothers who smoked only during pregnancy are expected to have babies which are about 160 grams lighter. Mothers who smoked in the three months leading up to pregnancy as well are expected to have babies which are about 200 grams lighter. Finally, even mothers who only smoked leading up to pregnancy, but not during, also are expected to have lighter babies, but only by about 10 grams.

**Mother's Age** - Our model indicates a small but statistically significant positive effect of mother's age on birthweight. For every year a mother gets older, her baby is expected to be about 5 grams heavier.

**Race** - Race is expected to have a significant impact on birthweight. However, this effect is made hard to interpret because of the interaction term. This interaction and effect will be explored more below in the interaction section.

**Infant Mortality Rates** - Infant mortality rates are used as a proxy for measuring socioeconomic status, and is expected to have a significant impact on birthweight. However, this effect is made hard to interpret because of the interaction term. This interaction and effect will be explored more below in the interaction section.

**Interaction Terms Show The Following** - An "other non-white" mother has her child's birth weight more negatively affected by an increase in mortality rate (birth weight drops by ~255 grams for each percentage of increase in mortality) than is the case for other races of mothers (children of mothers of other races only drop by ~120 grams). The MRACER Other Non-White:mortality is -255, whereas the coefficients for the other race:mortality interaction terms are around -120.

### Results from Quantile Regression and Model Coefficients

**Quantile Regression** The quantile regression reveals the following:

1. Only Smoking Before Pregnancy: For the lightest babies (at the 0.05 quantile), smoking only before pregnancy led to a 11.5 gram drop in birth weight. At the median, this effect was -2.7 grams. For the largest babies, at the 95th quantile, this effect was +55.9 grams.

2. Only Smoking During Pregnancy: For the lightest babies (at the 0.05 quantile), smoking only during pregnancy led to a 115.3 gram drop in birth weight. At the median, this effect was -104.5 grams. For the largest babies, at the 95th quantile, this effect was -55.9 grams.

3. Smoking Before and During Pregnancy: For the lightest babies (at the 0.05 quantile), smoking before and during pregnancy led to a 203.4 gram drop in birth weight. At the median, this effect was -198.1 grams. For the largest babies, at the 95th quantile, this effect was -169.4 grams.

### Conclusion

In conclusion, we have answered the 4 main questions of this case study:

1. The relevant factors are Sex of Child, Gestational period, parity, plurality, smoking, mother's age, mother's race, and mortality. There was also an interaction effect between mortality and mother's race.

2. The effects are explored above. Some specific ones include: 1. As gestational period increases, birthweight increases 2.Twins and triplets (or more) are expected to be much lighter than single-birth babies, by about 310-320 grams 3.Smoking is expected to have a negative effect on babies' birthweights. Smoking of any kind decreases birth weight, and smoking only before pregnancy leads to a 200 gram decrease in birthweight

3. We ended up using robust regression on model4 to attempt to solve the problem of heteroscedasticity, and when that did not work well enough, we used quantile regression. Model4 ended having the lowest MSE, even lower than the robust model, after examination with cross validation. Otherwise, there was a little bit of a departure from normality.

4. Smoking is expected to have a negative effect on babies' birthweights. Mothers who smoked only during pregnancy are expected to have babies which are about 160 grams lighter. Mothers who smoked in the three months leading up to pregnancy as well are expected to have babies which are about 200 grams lighter. Finally, even mothers who only smoked leading up to pregnancy, but not during, also are expected to have lighter babies, but only by about 10 grams.